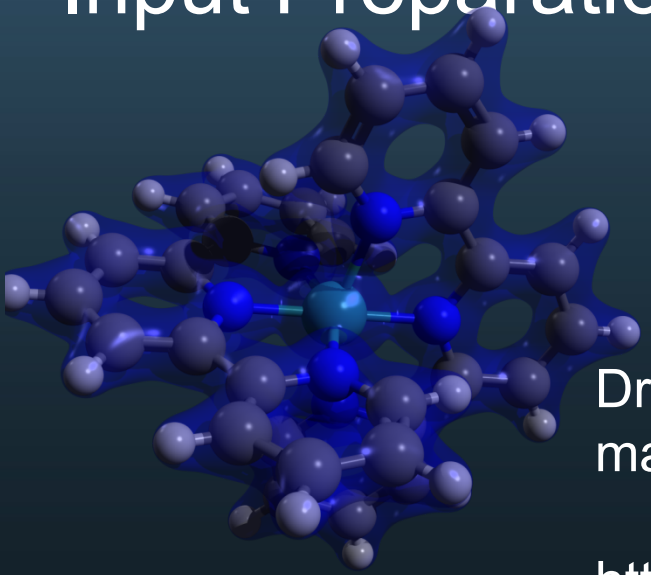
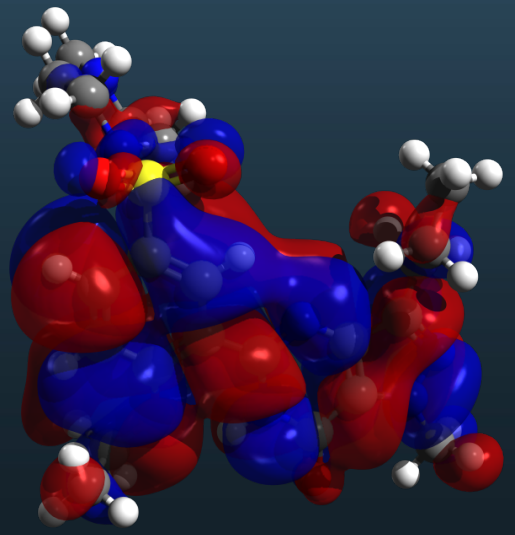
Input Preparation, Data Visualization & Analysis



June 8, 2013
LA-SiGMA
Baton Rouge, LA

Dr. Marcus D. Hanwell
marcus.hanwell@kitware.com

http://openchemistry.org/

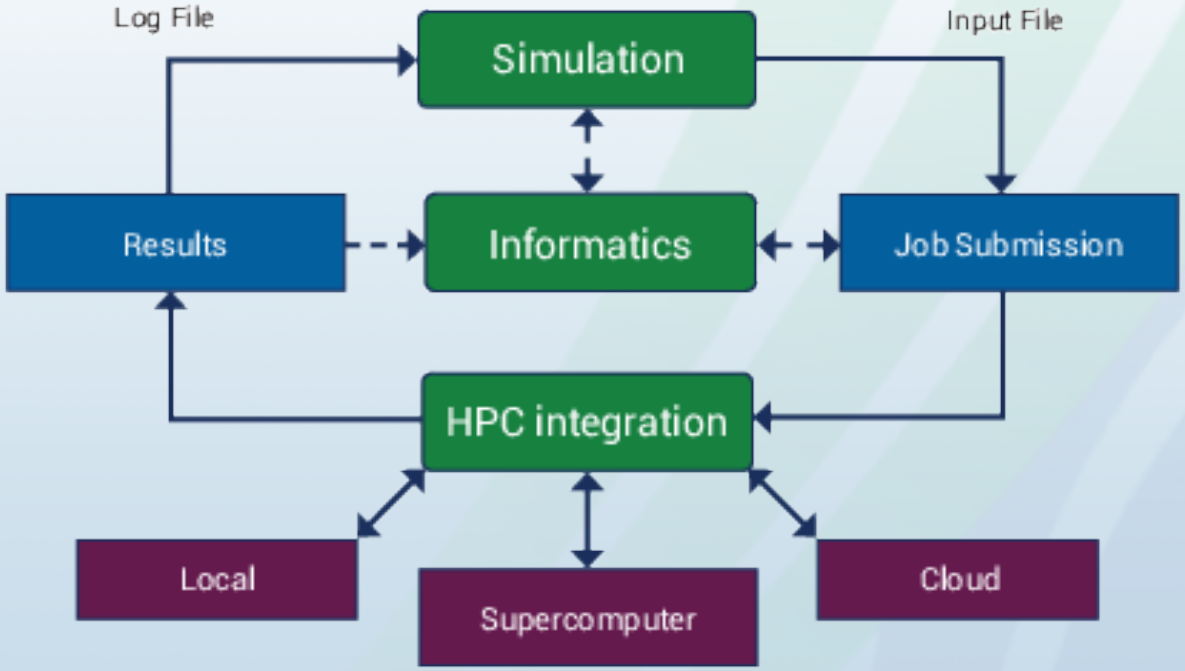


Outline

- **Introduction**
- Kitware
- Open Chemistry
  - Avogadro 2
  - MoleQueue
  - MongoChem
- The Future
- Summary

Introduction

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



Kitware

Introduction

- Bringing real change to chemistry
  - Open-source frameworks
  - Developed openly
  - Cross-platform compatibility
  - Tested and verified
  - Contribution model
  - Supported by Kitware experts
- Liberally-licensed to facilitate research

Open Chemistry Development Team

- Inter-disciplinary team at Kitware
- The first three worked on open-source chemistry in their spare time
- The final two are computer scientists with years of open-source experience
- Seeking partners in industry & research, labs

Kitware

Outline

- Introduction
- **Kitware**
- Open Chemistry
  - Avogadro 2
  - MoleQueue
  - MongoChem
- The Future
- Summary

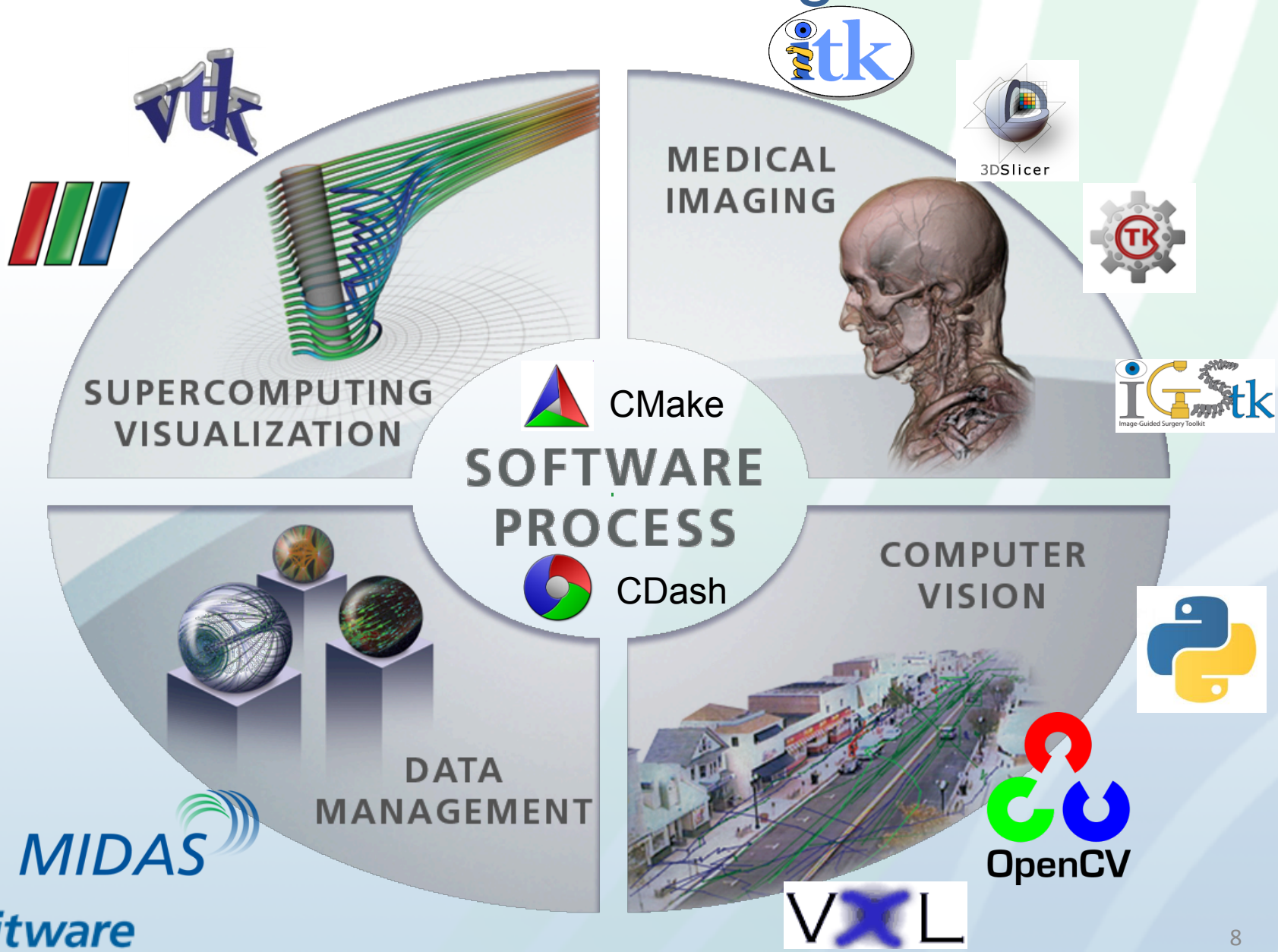
Kitware

- Founded in 1998 by five former GE Research employees
- 118 current employees; 39 with PhDs
- Privately held, profitable from creation, no debt
- Rapidly Growing: >30% in 2011, 7M web-visitors/quarter
- Offices
  - Clifton Park, 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

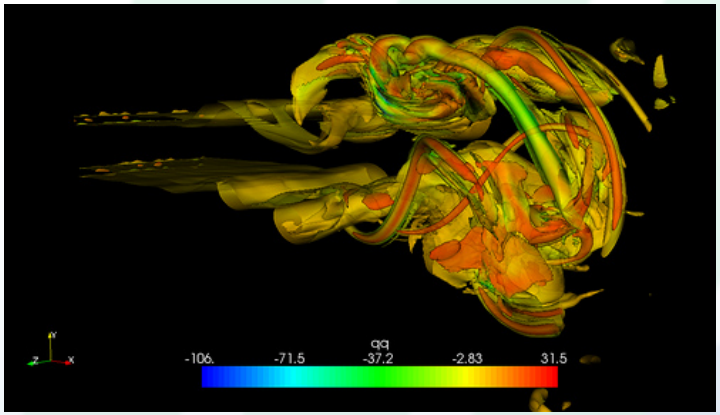
Kitware

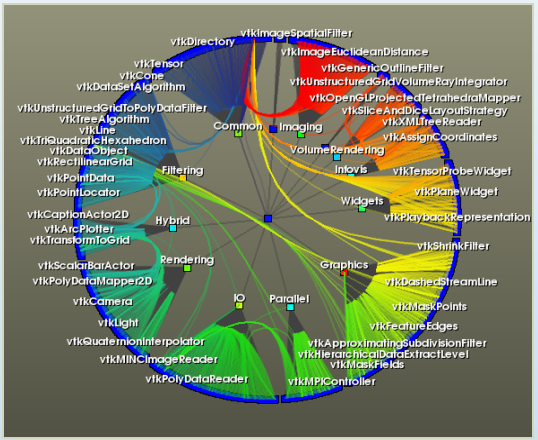
Kitware: Core Technologies

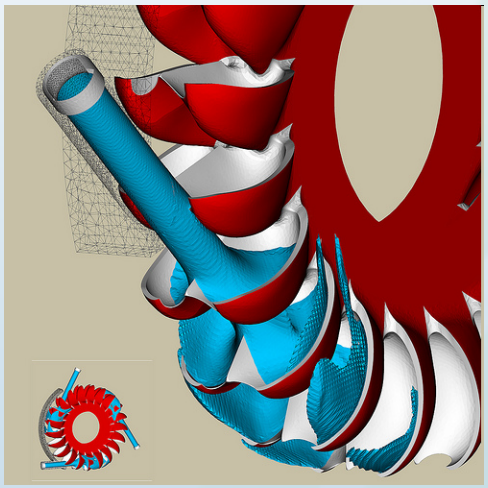


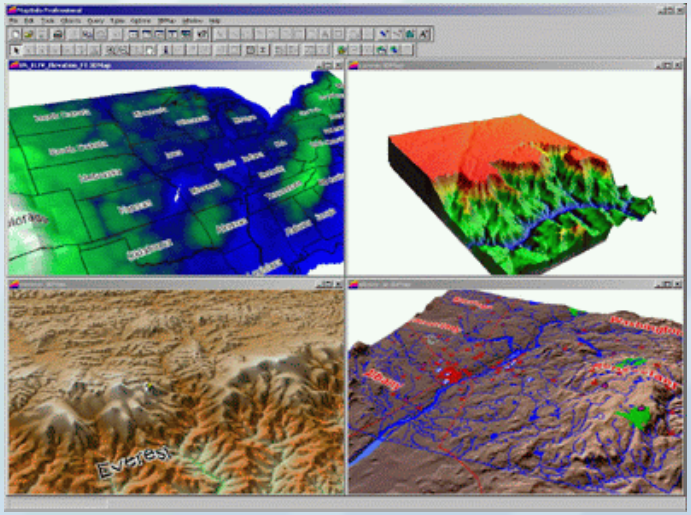
Supercomputing Visualization

- Scientific Visualization
- Informatics
- Large Data Visualization
- 3D Interaction
- Volume Rendering





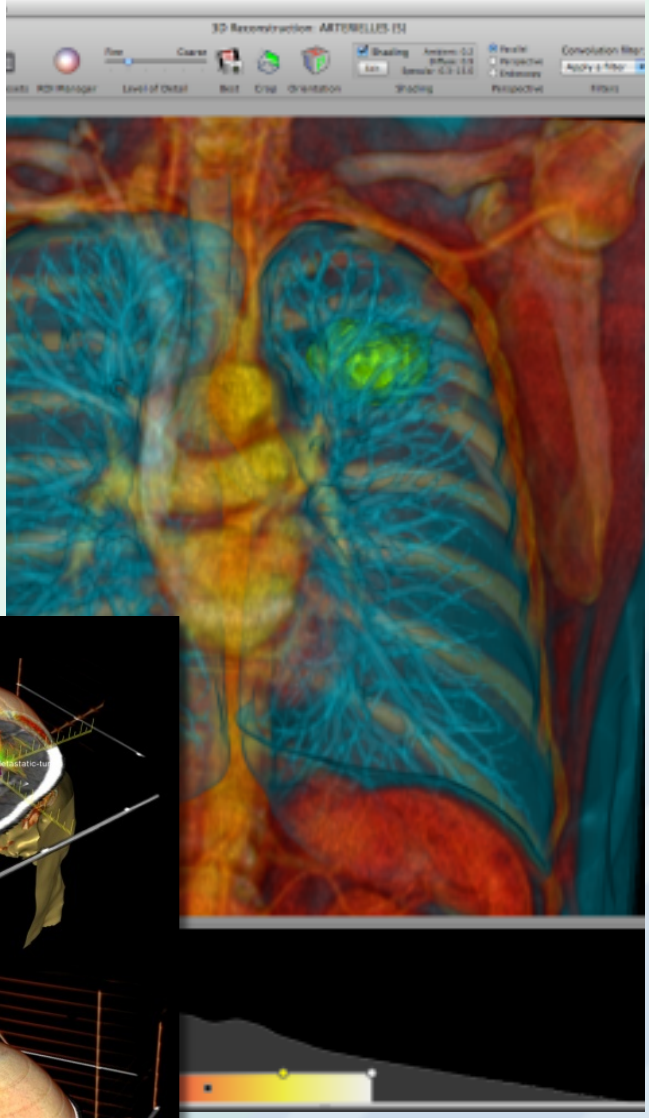


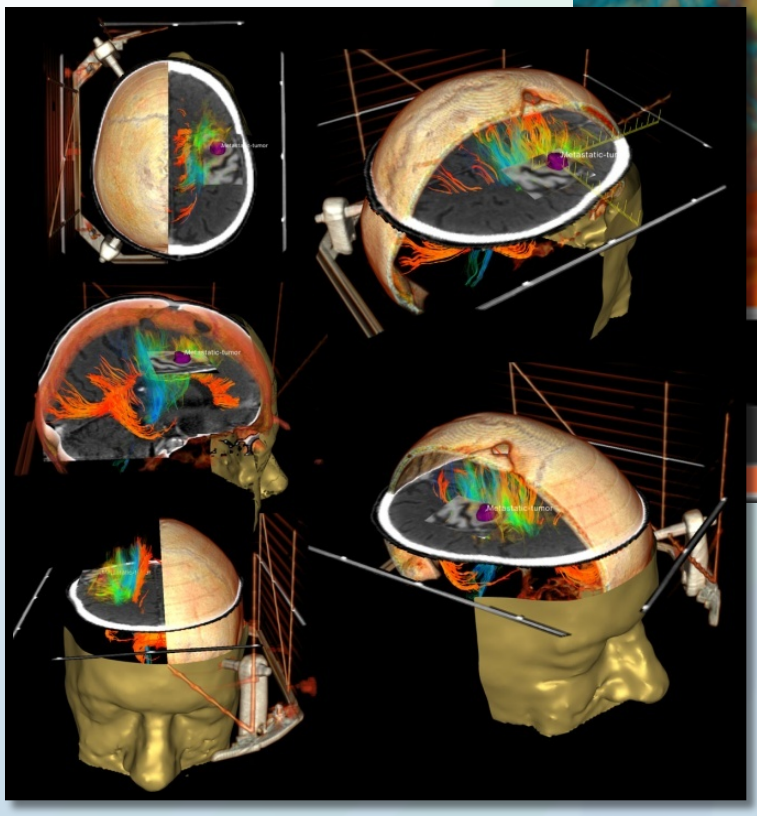


Kitware

Medical Image Analysis

- Image Processing
- Segmentation
- Registration
- Measurement & Analysis

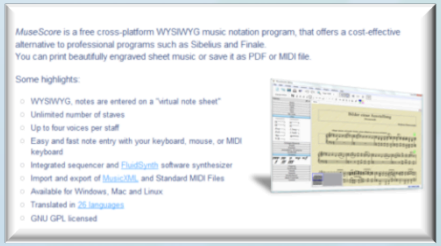




Kitware

CMake provides the software process for many popular projects

Allegro library
Armadillo
Avidemux
Awesome
Blender 3D
Bullet Physics Engine
Chicken Scheme
Chipmunk physics engine
Clang
Compiz
Conky
Doomsday Engine
Drishti
Gammu
GDCM
Gmsh
Hypertable
Hugin
iCub
IGSTK
ITK
KDE SC 4
Kicad
LMMS
LLVM
MariaDB
MiKTeX
MuseScore
MySQL
OGRE
OpenSceneGraph
OpenSync
OpenCV
ParaView
Poppler
PvPGN
Quantum GIS
QutIM
Raw Therapee
ROS
Scribus
Second Life
Spring RTS
SuperTux
Slicer
Stellarium
VTK
VXL
YARP



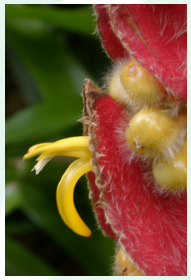
Kitware

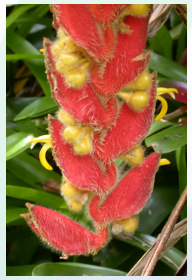
ARBOR
revolutionary workflows

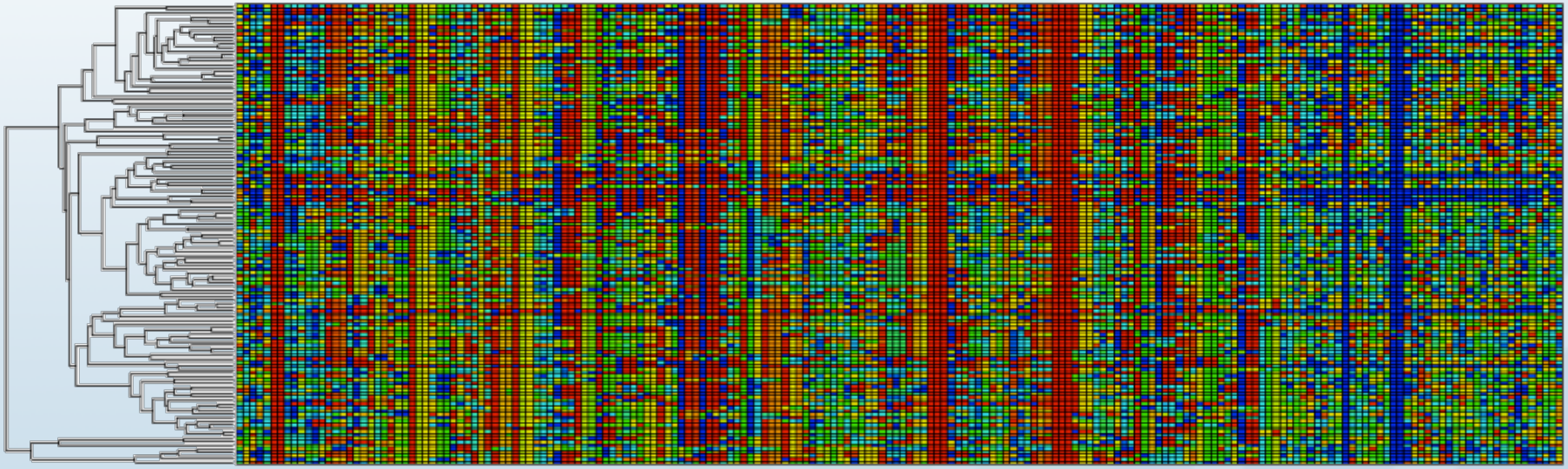
Arbor is an NSF funded project to enable evolutionary biological research by making it easy for biologists to

- Create
- Test
- Visualize

Algorithms on the Tree of Life. Below is the evolutionary tree for *Heliconia* (Lobster Claw) plants coupled to a character matrix of observational data such as color, feature measurements and range.



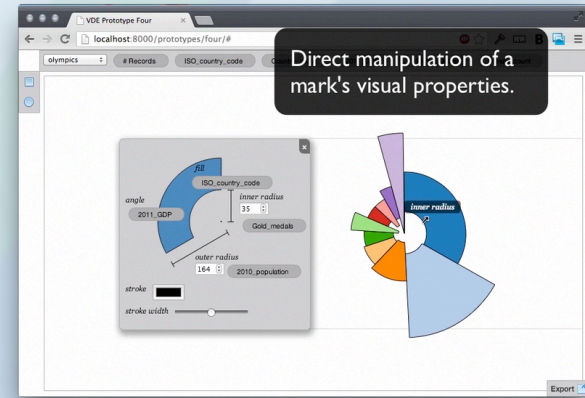




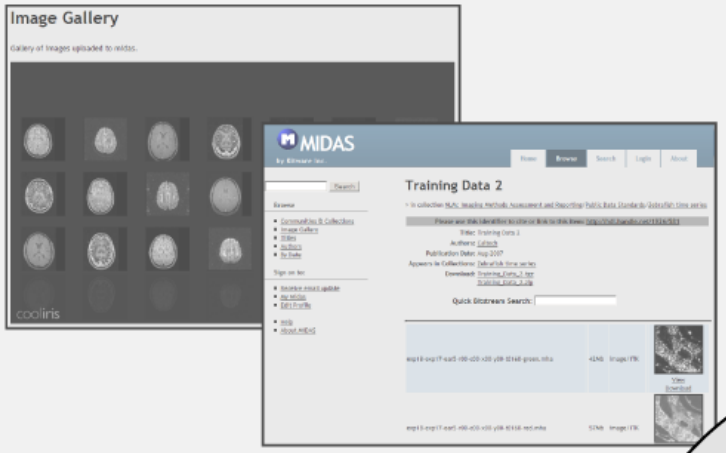
Kitware

DARPA XData Project

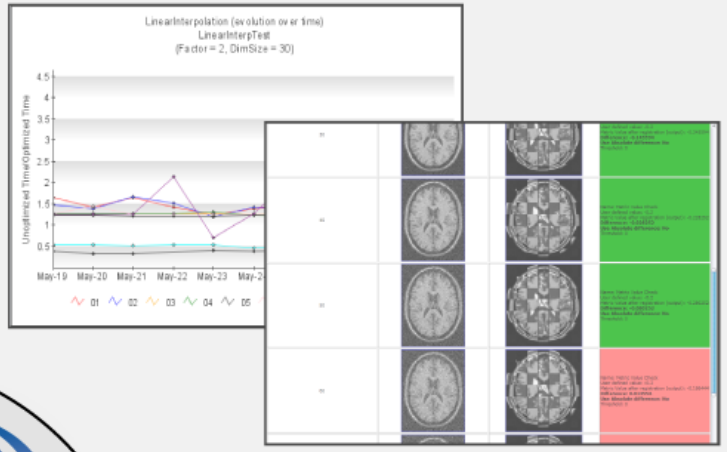
- Addressing needs of big data analysis
- Large, collaborative project
- PI: Jeffrey Baumes, Kitware Inc.
  - Jeffrey Heer, Stanford
  - Hanspeter Pfister, Harvard
  - John Stasko, Georgia Institute of Technology
  - Miriah Meyer, University of Utah
  - Curtis Lisle, KnowledgeVis LLC



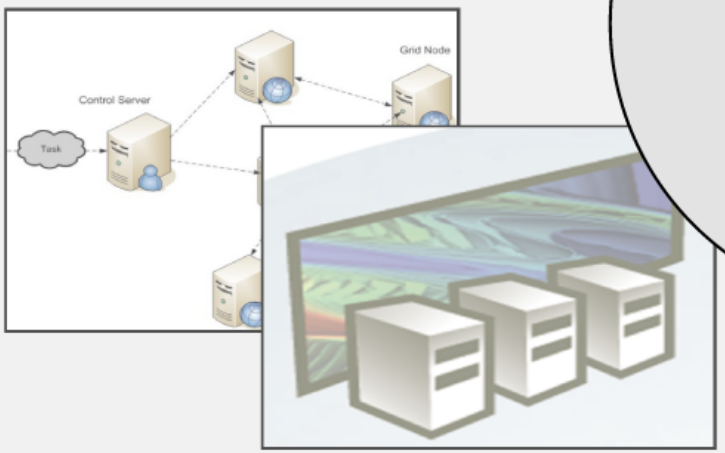
Kitware



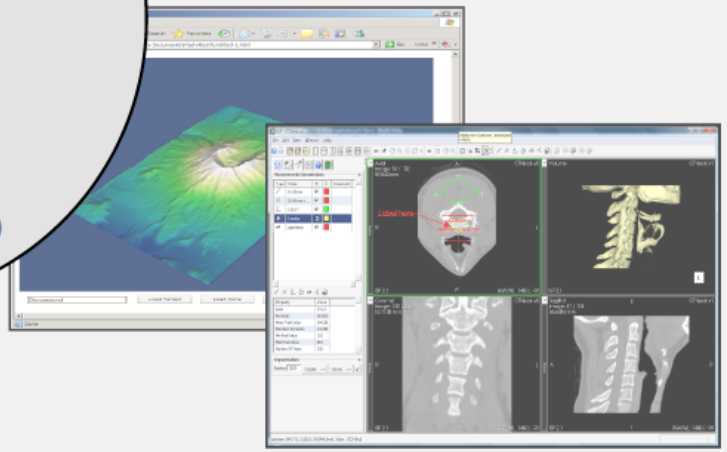
Digital Storage



Online Reporting



Server-Side Processing

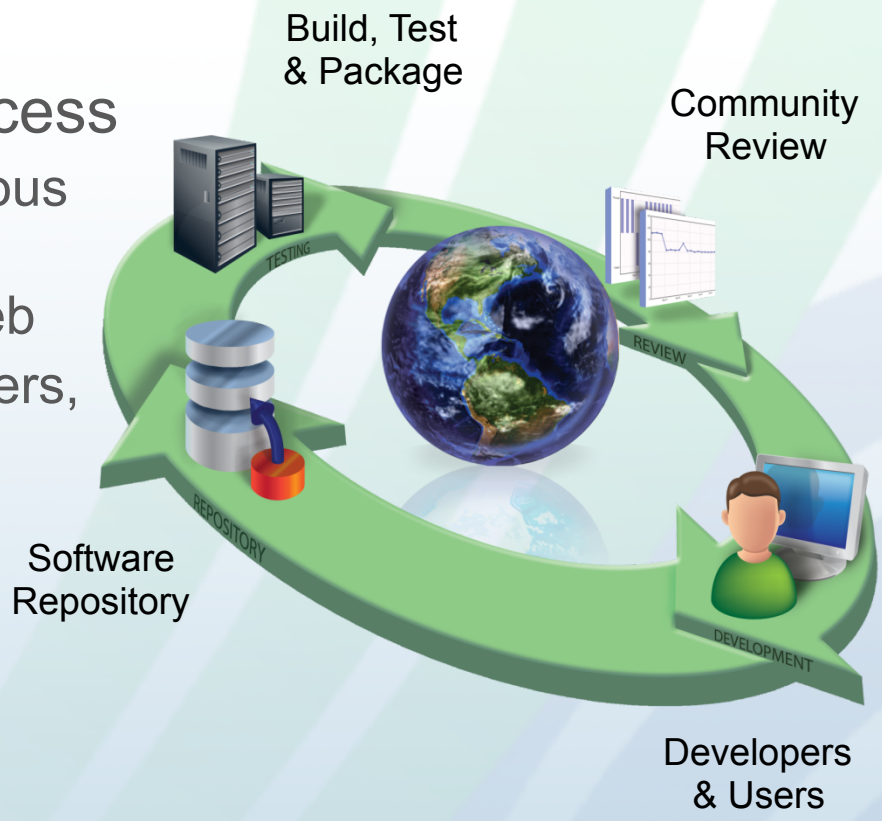


Interactive Visualization

Kitware

Building Community

- Communities are grown around open source projects
- Using Kitware software process
  - Ensuring quality with continuous testing
  - Code contributions via the web
  - Public mailing lists, bug trackers, and code review
- Promoting projects and participation
  - Publications
  - Conferences
  - Workshops



Business Model: Open Source

- Open-source Software
  - Normally BSD-licensed
  - Collaboration platforms
- Collaborative Research and Development
- Technology Integration
- Service and Support
- Consulting
- Training and webinars

Business Model: Open Source

- Open-source platforms used in:
  - Research
  - Teaching
  - Commercial applications
- Software is created by:
  - Internationally-recognized (Kitware) experts
  - Extended open-source communities
- Using a rigorous, quality-inducing software development process

Commercialization Strategy

- Services & Consulting Model
  - Kitware develops widely-used software frameworks and serves them through consulting.
    - Collaborative R&D
    - Custom solution development
    - Value-added products (e.g., training, support, books)
    - Services comprise approximately 2/3 of the global software market
    - Companies such as IBM, HP, and Oracle realize massive business from services

Value of Open Source

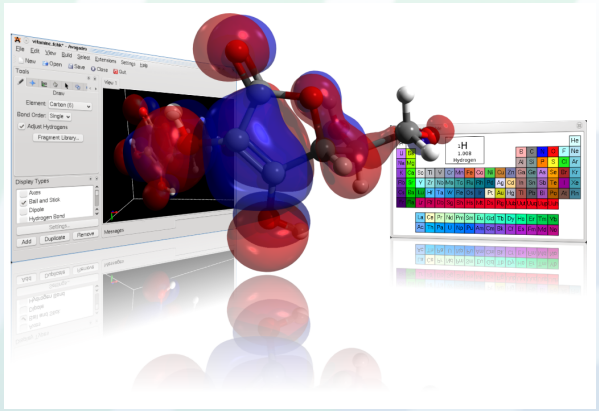
- Access to and ownership of the code
- Collaborative relationships are natural
- Rapid, responsive development process
- Partners can participate in development
- Reduced or (often) no licensing fees
  - Maintenance burden taken up by broader community
  - Often represents the greatest part of the cost of software

Outline

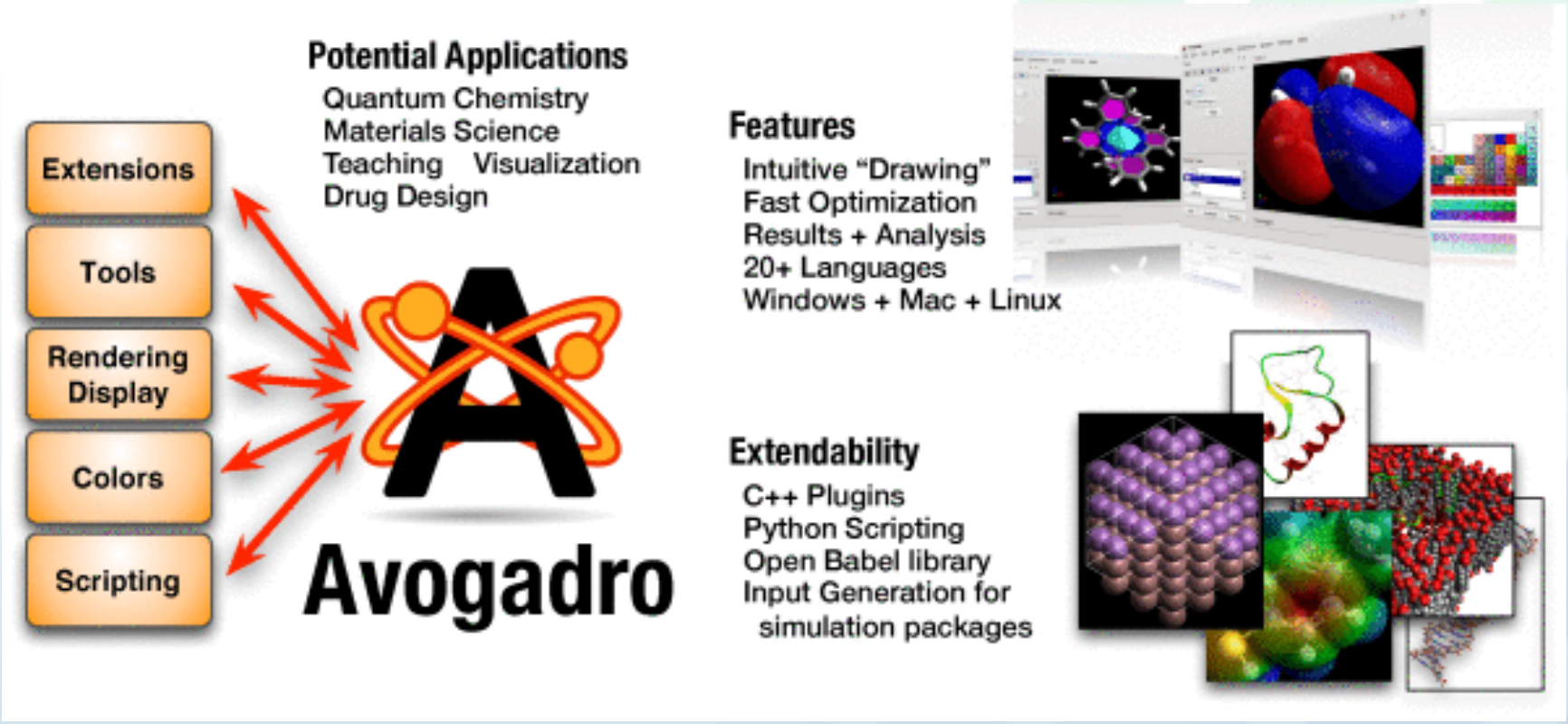
- Introduction
- Kitware
- **Open Chemistry**
  - Avogadro 2
  - MoleQueue
  - MongoChem
- The Future
- Summary

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, a KDE educational tool
- Over 300,000 downloads, 20+ translations



Avogadro Paper Published 8/13/12



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

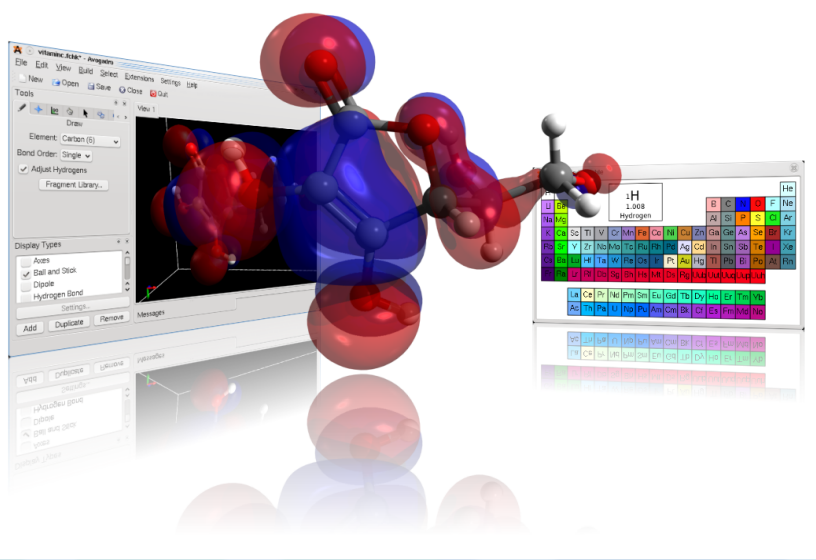
NWChem, FoX, Avogadro Paper Published 5/24/13

http://www.jcheminf.com/content/5/1/25

Kitware

Vision

- Advancing the state-of-the-art
- Tight integration is needed
  - Computational codes
  - Clusters/supercomputers
  - Data repositories
  - Reduce, reuse, and 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 for easily runing codes
  - Cheminformatics to store, index, and analyze
- Each program 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

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

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

OpenChemistry.org

- Web presence to promote Open Chemistry
- Hosting of project-specific pages
- Providing an identity for related projects
- Promote shared ownership of projects
  - Website
  - Code submission and review
  - Testing infrastructure
  - Wiki, mailing lists, news, and galleries

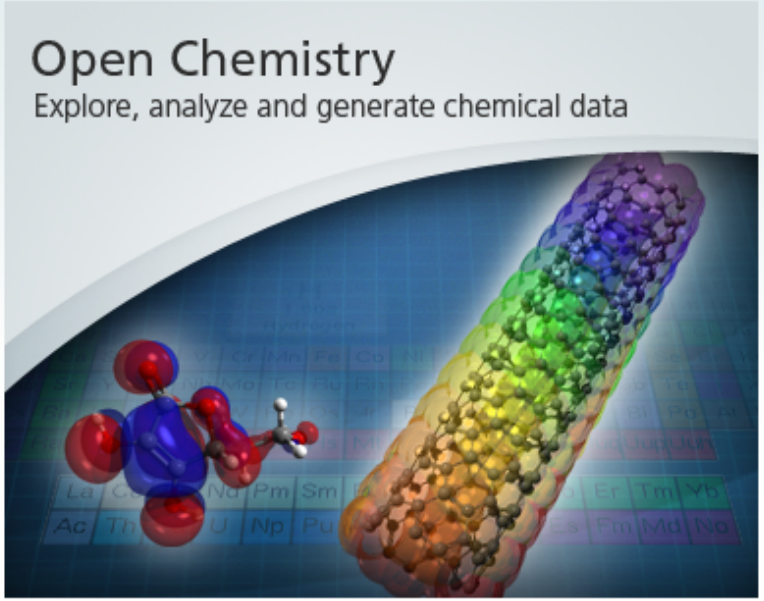
Open Chemistry

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 >

05.29.2013 New Open Access Article on a Collaborative Project Between NWChe...

08.17.2012 Avogadro Featured in Journal of Cheminformatics

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

Events

06.07.2013 LA-SIGMA

Blog Posts

05.22.2013 New Input Generator Framework in Avogadro 2

05.01.2013 Using VTK's Image Regression Tests in Avogadro 2

04.11.2013 First Open Chemistry Beta Release

Applications Being Developed

- Three independent applications
- Communication handled with local sockets

Avogadro 2: Structure editing, input generation, output viewing, and analysis

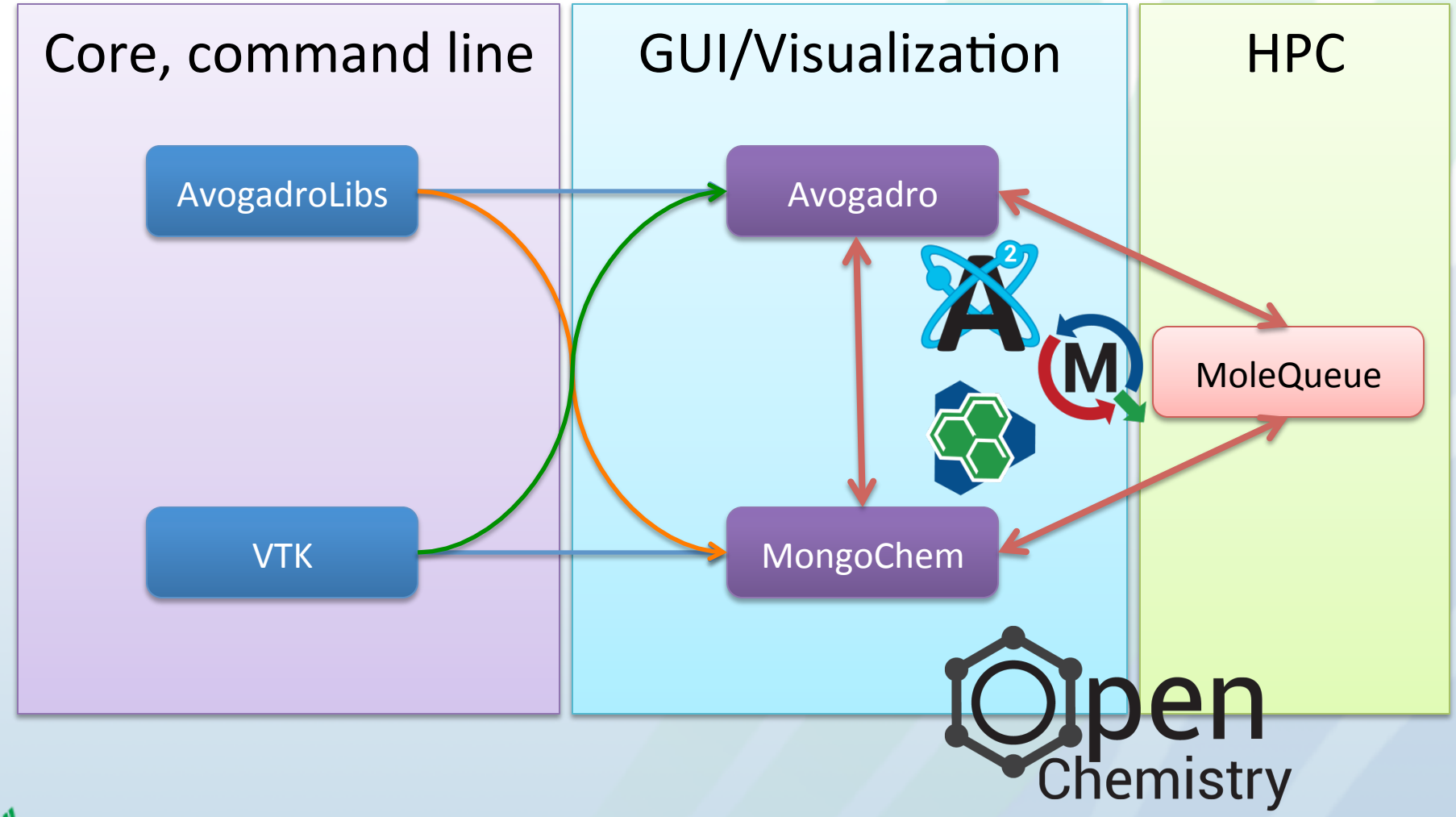
MoleQueue: Running local and remote jobs in standalone programs, and management

MongoChem: Storage of data, searching, entry, and annotation

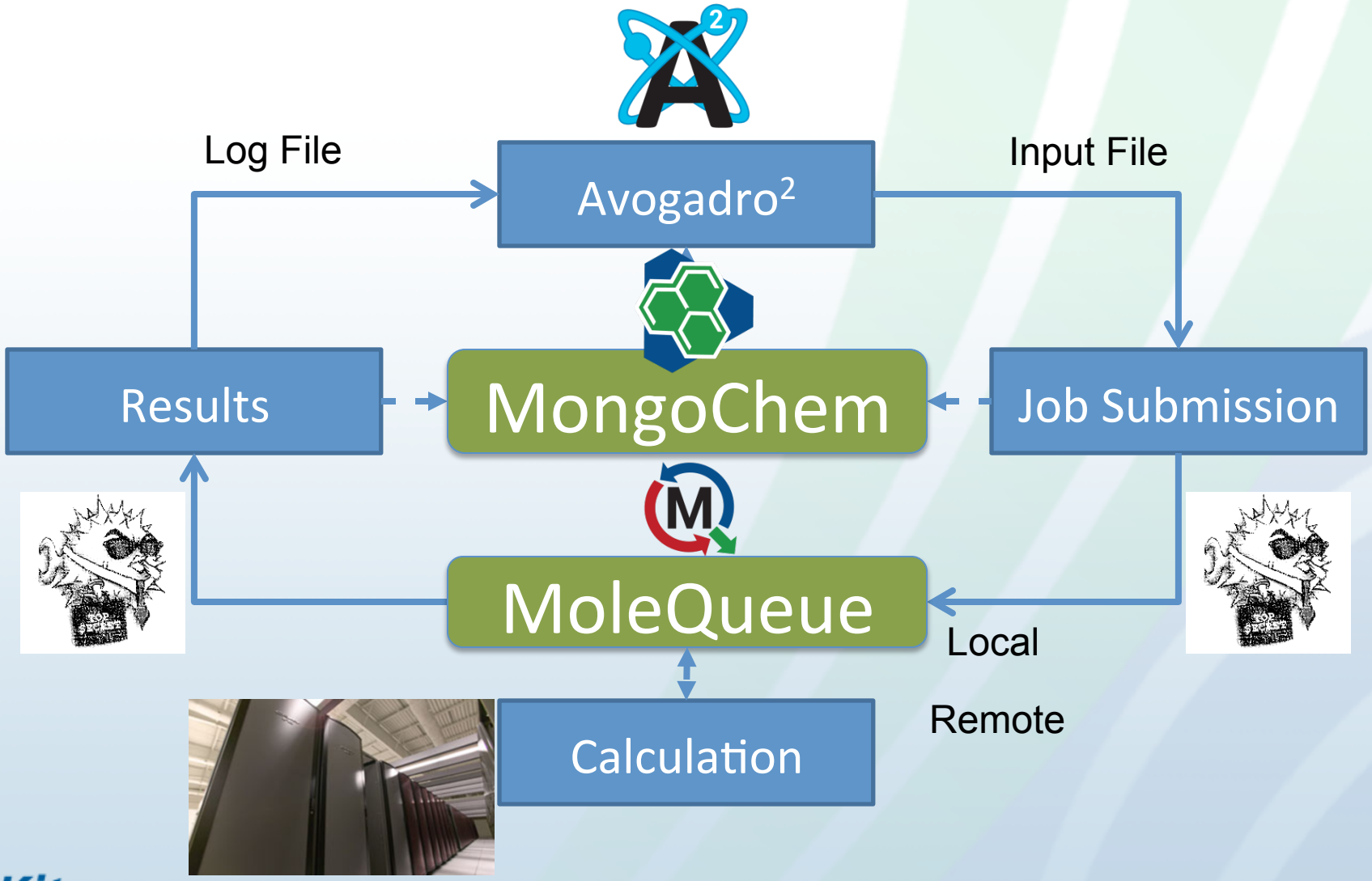
Open Frameworks

- 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: Chemistry visualization and data structures, use of above

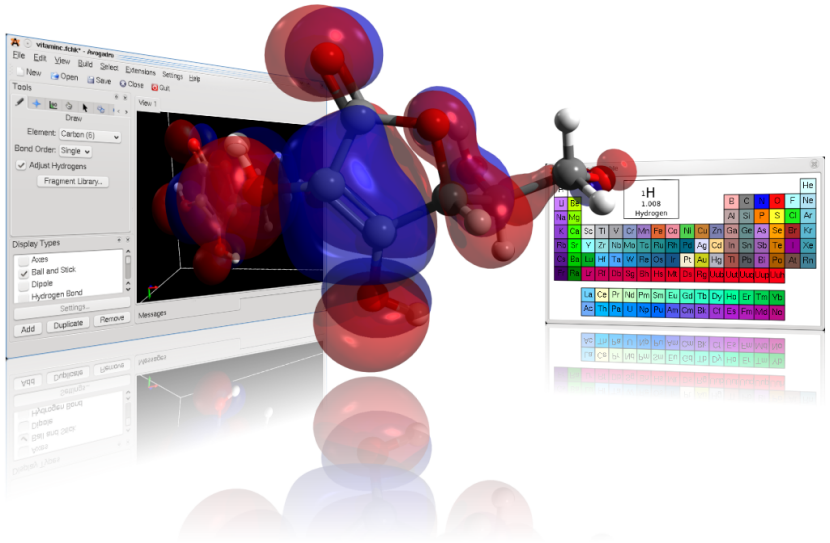
Project Diagram: Libraries/Apps



Workflow in Open Chemistry



Avogadro^2



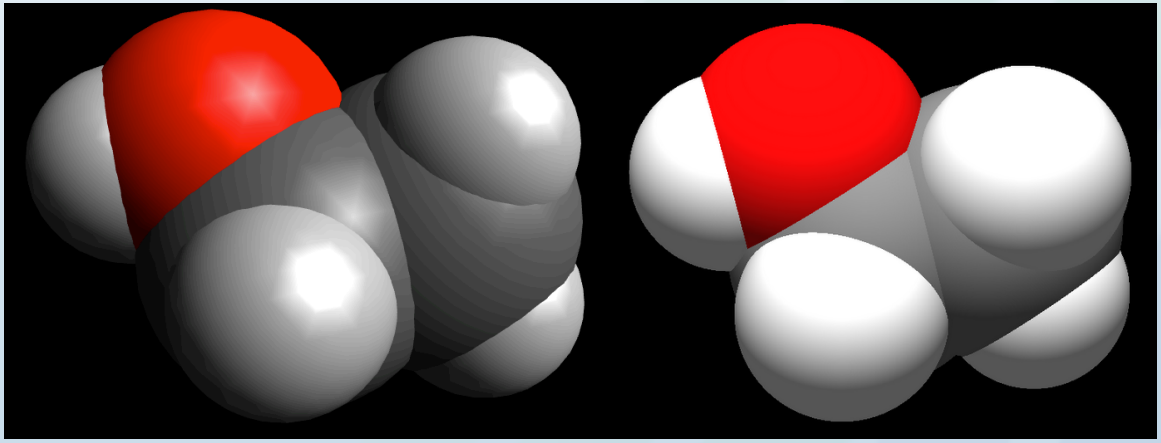
- Rewrite of Avogadro
- Split into libraries & application (plugin-based)
- Still one of very few open source **editors**
- Still using Qt, C++, Eigen, OpenGL, CMake
- Use AvogadroLibs for core data
- Introduces client-server dataflow/patterns
- New, efficient rendering code
- More liberally-licensed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

Advanced Impostor Rendering

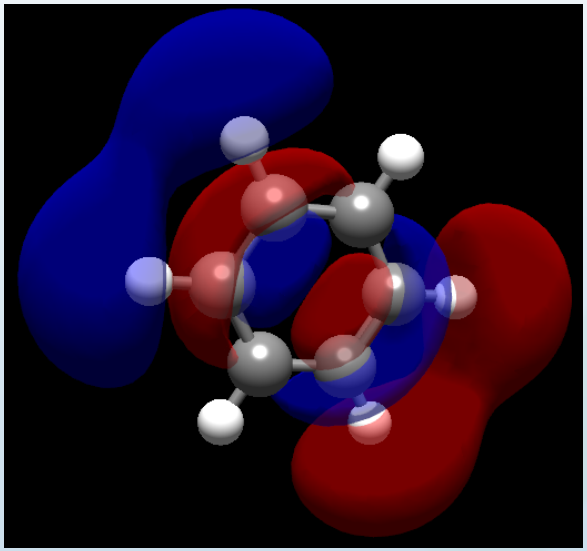
- Using a scene, vertex buffer objects, and OpenGL shading language
- Impostor techniques
  - Sphere goes from 100s of triangles to 2!
  - No artifacts from triangulation
  - Scales to millions of spheres on modest GPU

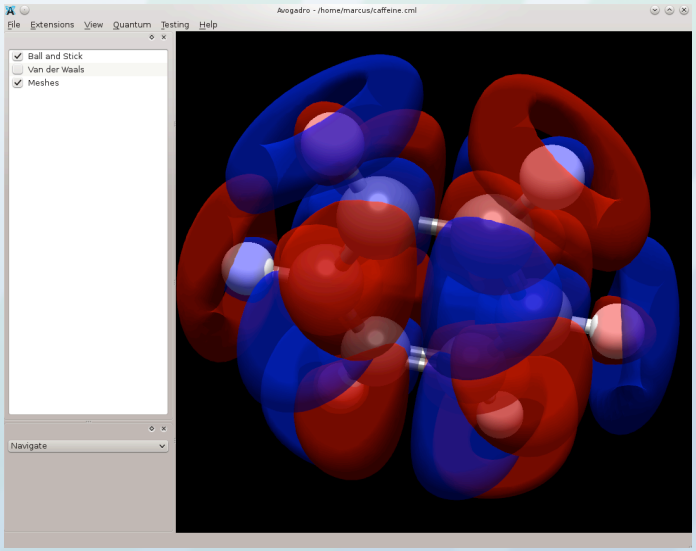


Kitware

Electronic Structure Visualization

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

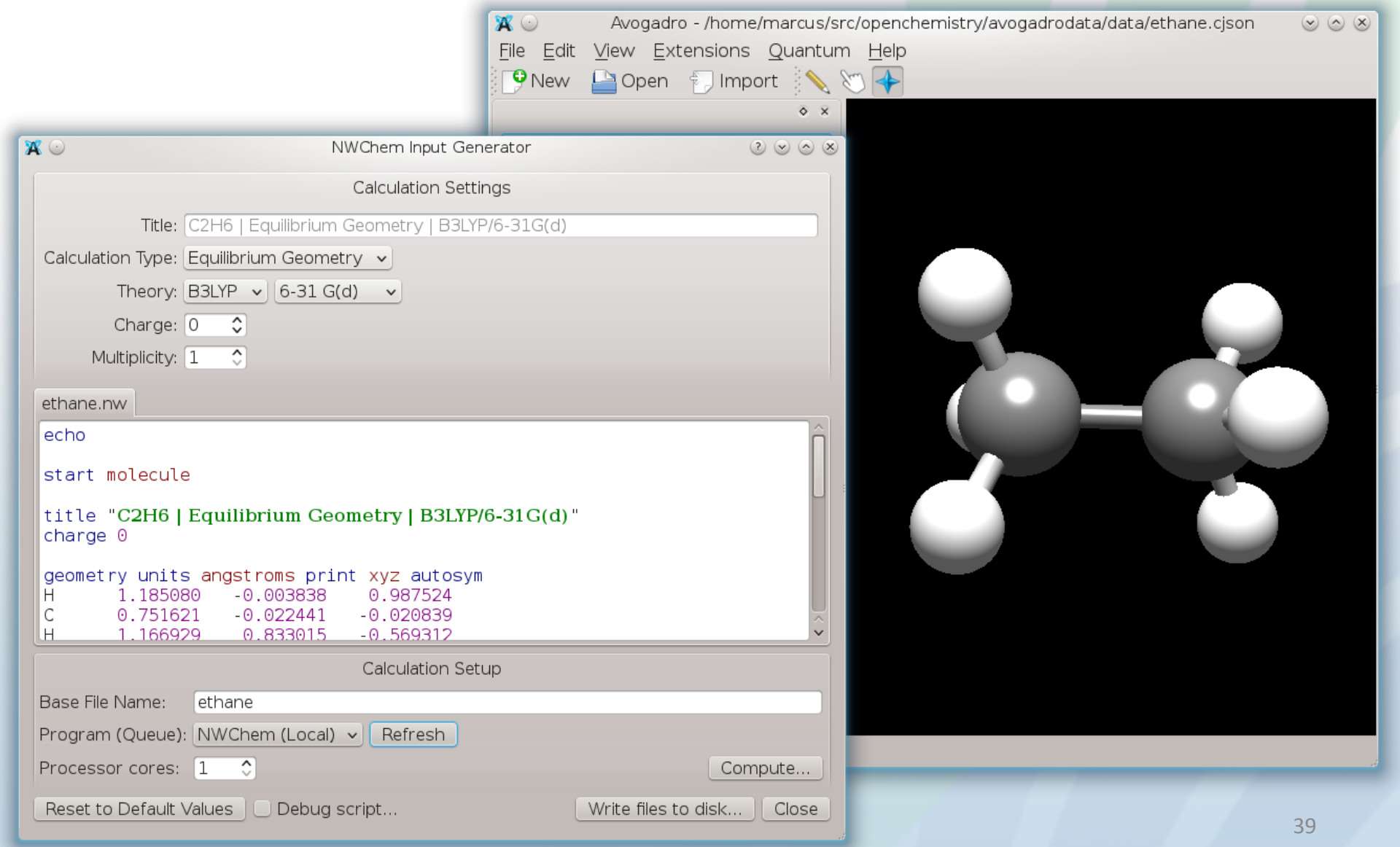




Scriptable Simulation Input Generator

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

Avogadro to Input Deck



Quantum Data in AvogadroLibs

- Reads in key quantum data
  - Basis set used in calculation
  - Eigenvectors for molecular orbitals
  - Density matrix for electron density
  - Standard geometry
- Multi-threaded calculation
  - Produces regular grids of scalar data
  - Molecular orbitals, electron density…

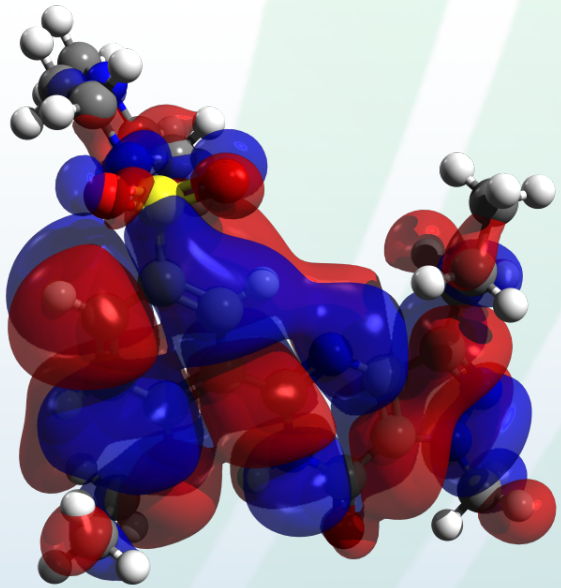
Molecular Orbitals and Electron Density

- 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

Calling Stand-alone Programs

- Many are already supported:
  - NWChem, GAMESS, GAMESS-UK, Molpro, Q-Chem, MOPAC, Gaussian, Dalton
  - Very easy to add more
- MongoChem and Avogadro 2 use libraries
  - Custom applications are simple
  - Now with simpler BSD licensing, testing, …
- Started looking at/prototyping other areas
  - Molecular dynamics, plane-wave, APBS

New CML I/O

- Development of modular CML code
- Allow for multi-pass parsing of CML
- Keep the CML closer to application
- Much faster, easier to extend and change
- Moving from simple CML to full semantic documents that can be edited
- Learned from previous work in VTK and Open Babel

File Format: CML & HDF5

- Leverage our experience with XDMF
- Early prototype already implemented
- CML stores semantic data
  - Name, formula, atoms, bonds
  - Computational code, theory, basis set
- HDF5 used to store heavy data
  - Basis set, intermediate data
  - Eigenvectors, SCF matrix
  - Volumetric data (MOs, electron density)

Avogadro: Client-Server

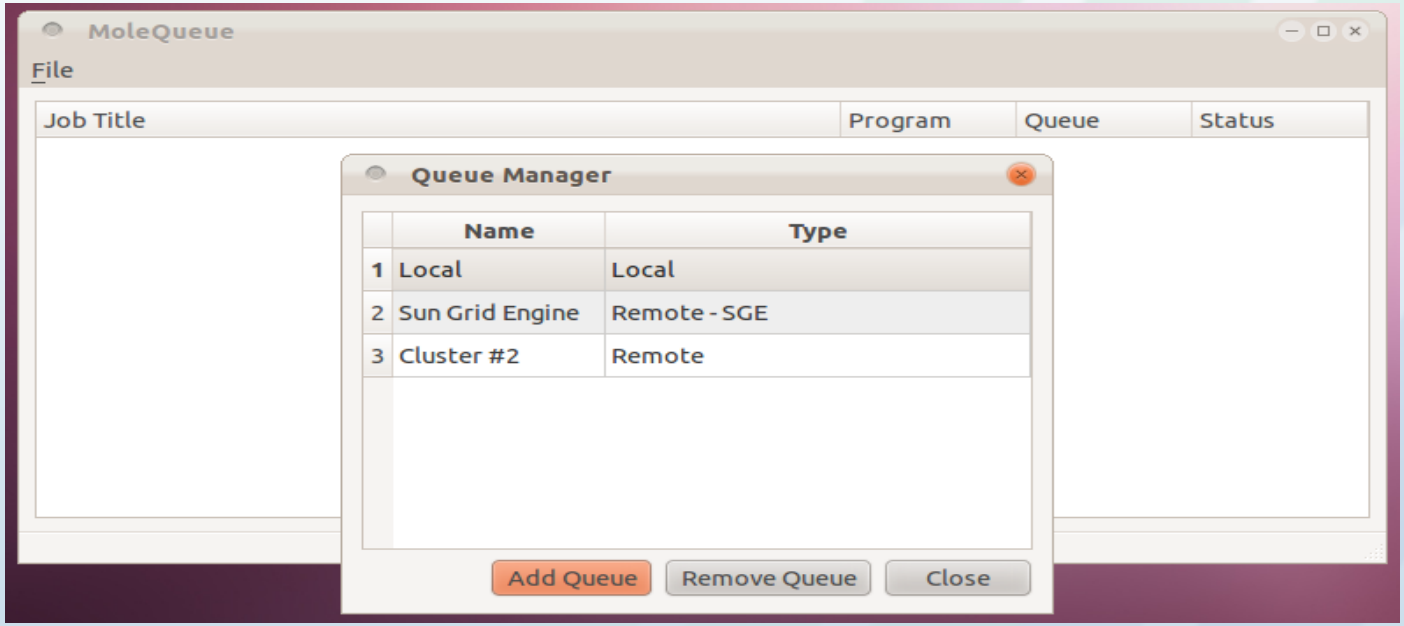
- Currently in early stages of development
- Off-loads more calculations to cluster
- Streams data, geometry
- Loading/creation of data remotely
- Analysis of large data
  - Processing nodes
  - Rendering nodes
- Scales to very large data

MoleQueue: Job Management

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

MoleQueue

- Supports configuration for a variety of remote clusters and queuing software
- Transparently switches between local and remote execution of codes



Kitware

MoleQueue: Queue Types

- Several transports implemented
  - Command line SSH/plink (Windows)
  - libssh2 (experimental)
  - HTTPS (SOAP)
- Several queue types
  - Local (execute and marshal processes)
  - Sun Grid Engine, PBS, SLURM
  - 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)

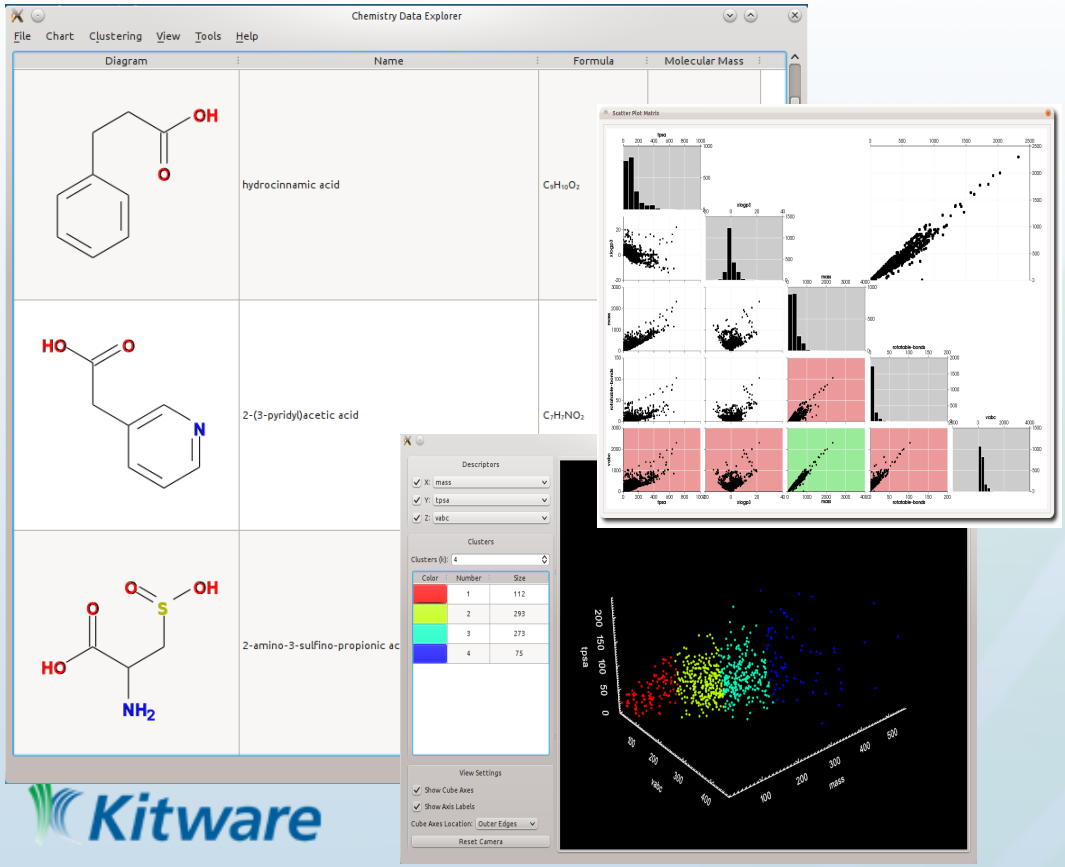

```
{
  "chemical json": 0,
  "name": "ethane",
  "inchi": "1/C2H6/c1-2/h1-2H3",
  "formula": "C 2 H 6",
  "atoms": {
    "elements": {
      "number": [ 1,  6,  1,  1,  6,  1,  1,  1 ]
    },
    "coords": {
      "3d": [  1.185080, -0.003838,  0.987524,
               0.751621, -0.022441, -0.020839,
               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
  }
}
```

MongoChem Overview

- A desktop cheminformatics tool
  - Chemical data exploration and analysis
  - Interactive, editable, and searchable database
- Leverages several open-source projects
  - Qt, VTK, MongoDB, Avogadro 2, Open Babel
- Designed to look at many molecules
- Spots patterns, outliers; runs many jobs
- Scales to studies with ~3 million structures

Architecture Overview

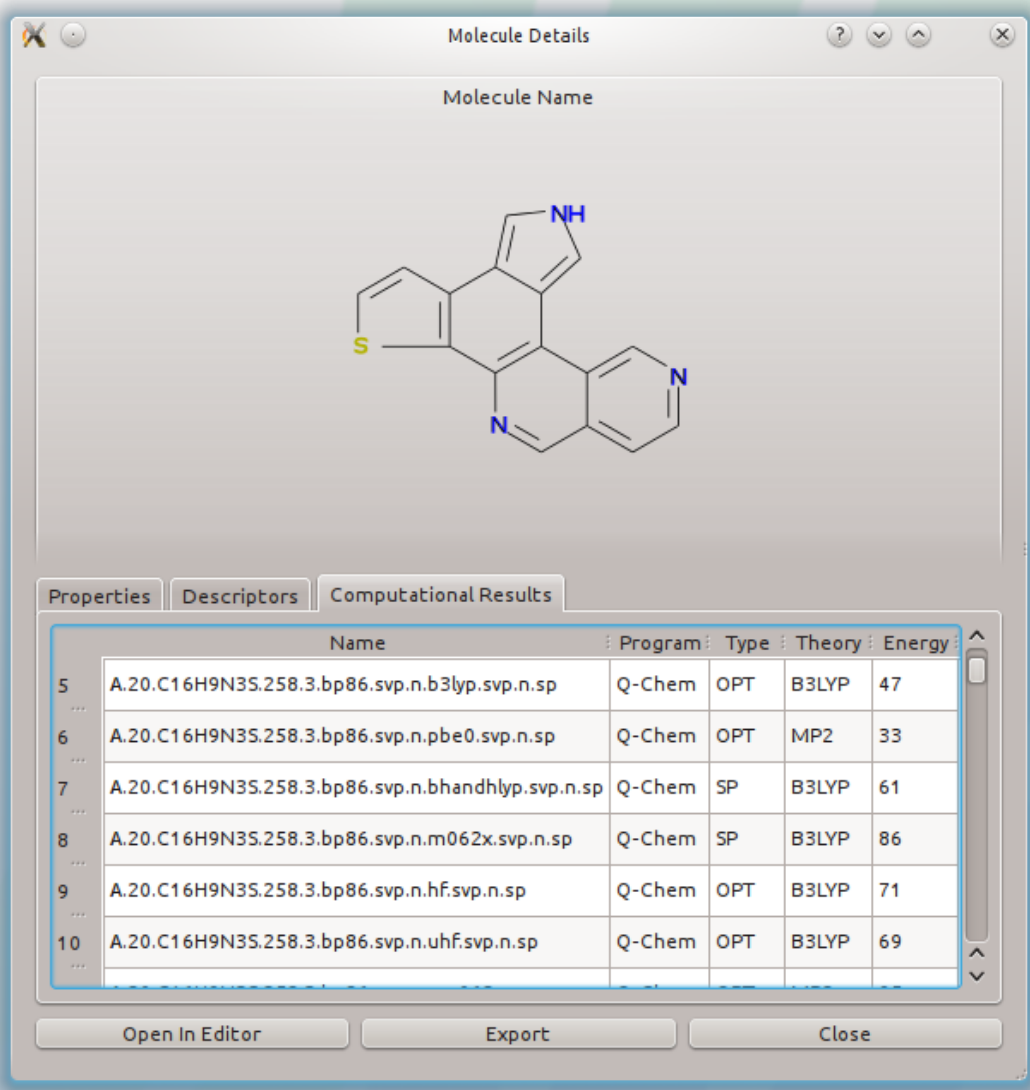
- Native, cross-platform C++ application built with Qt
- Stores chemical data in a NoSQL MongoDB database
- Uses VTK for 2D and 3D dataset visualization



Kitware

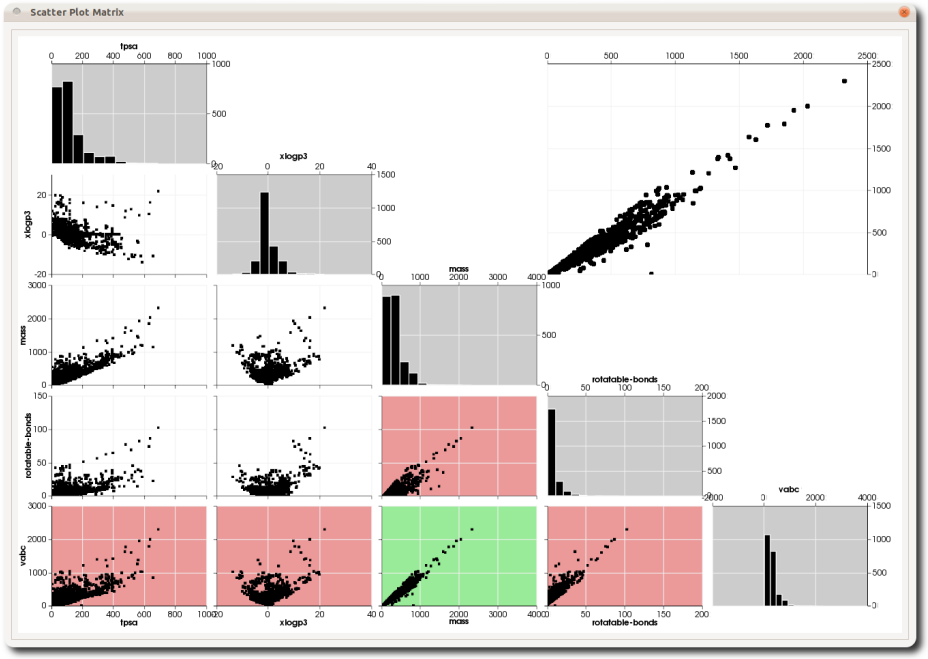
Computational Job Storage

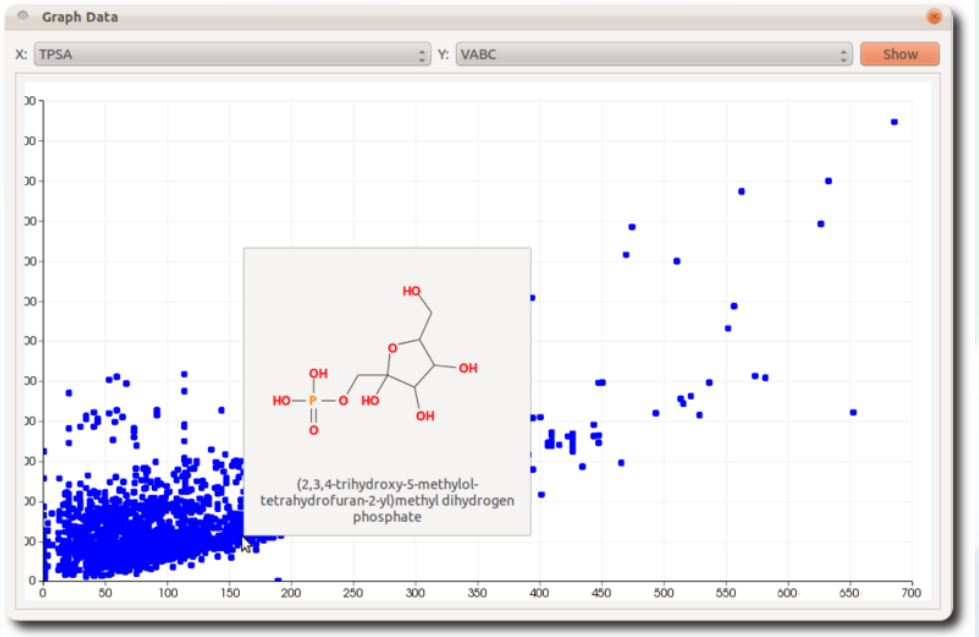
- Jobs associated with molecules
- Searchable based on structure/job parameters

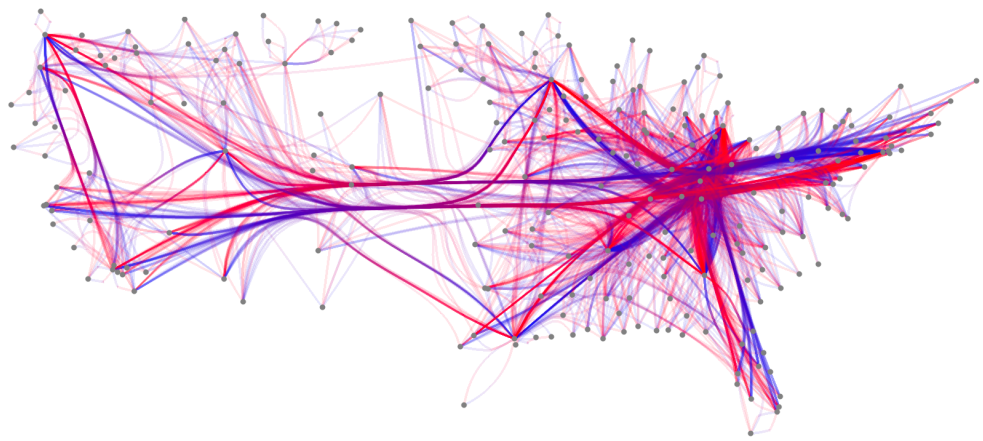


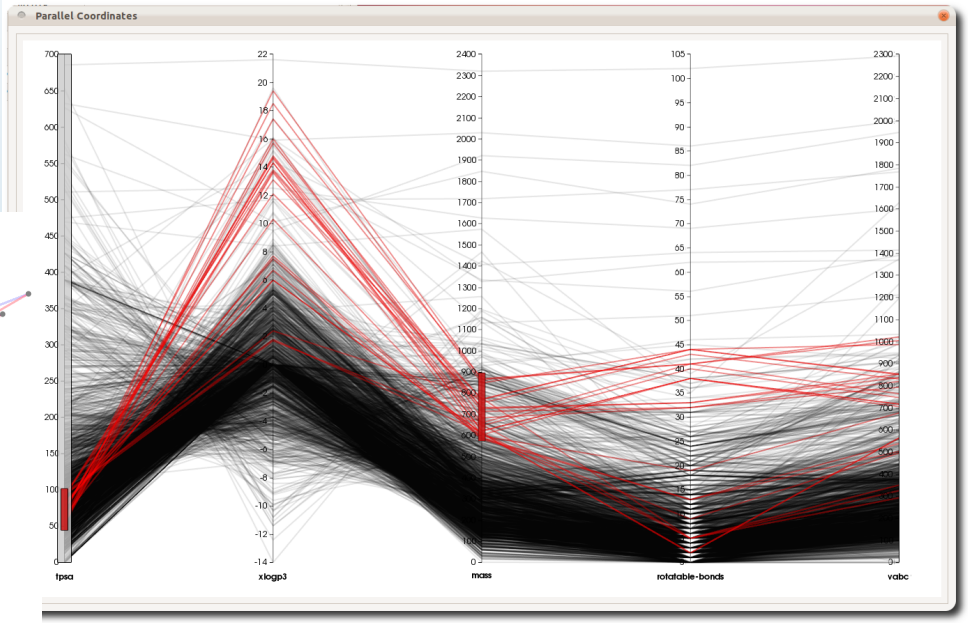
Kitware

Charts and Plots



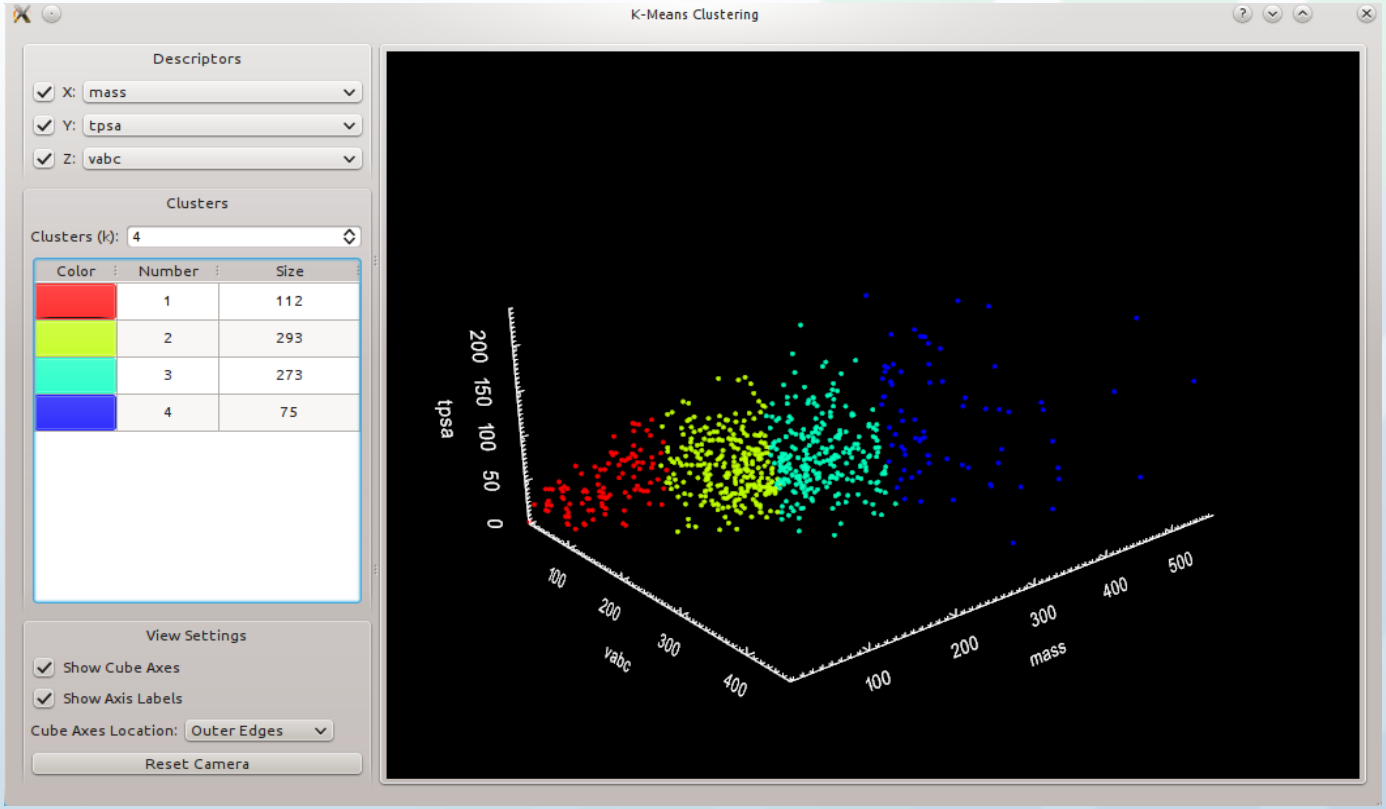






K-Means Clustering

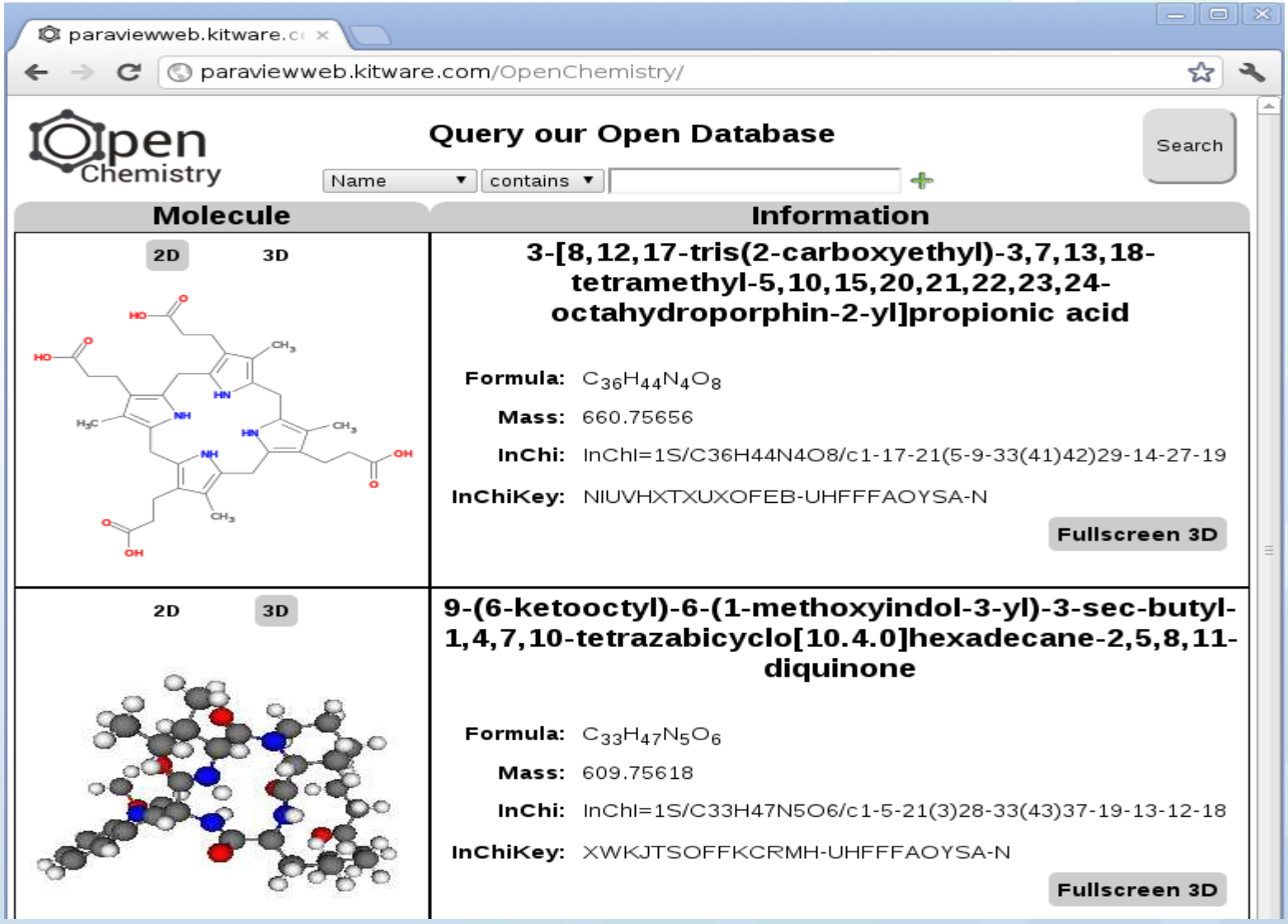
- >30 numerical molecular descriptors
- Extraction and filtering into clusters



ParaViewWeb and MongoChem

- Uses ParaView's client-server architecture
- Interactive 3D rendering
- Runs in any modern web browser
- Same MongoDB server as MongoChem
- Move more to the client JavaScript code
- Moving to a simple, Python-based server
  - Easy to add new APIs
  - Easy to deploy/integrate into other solutions

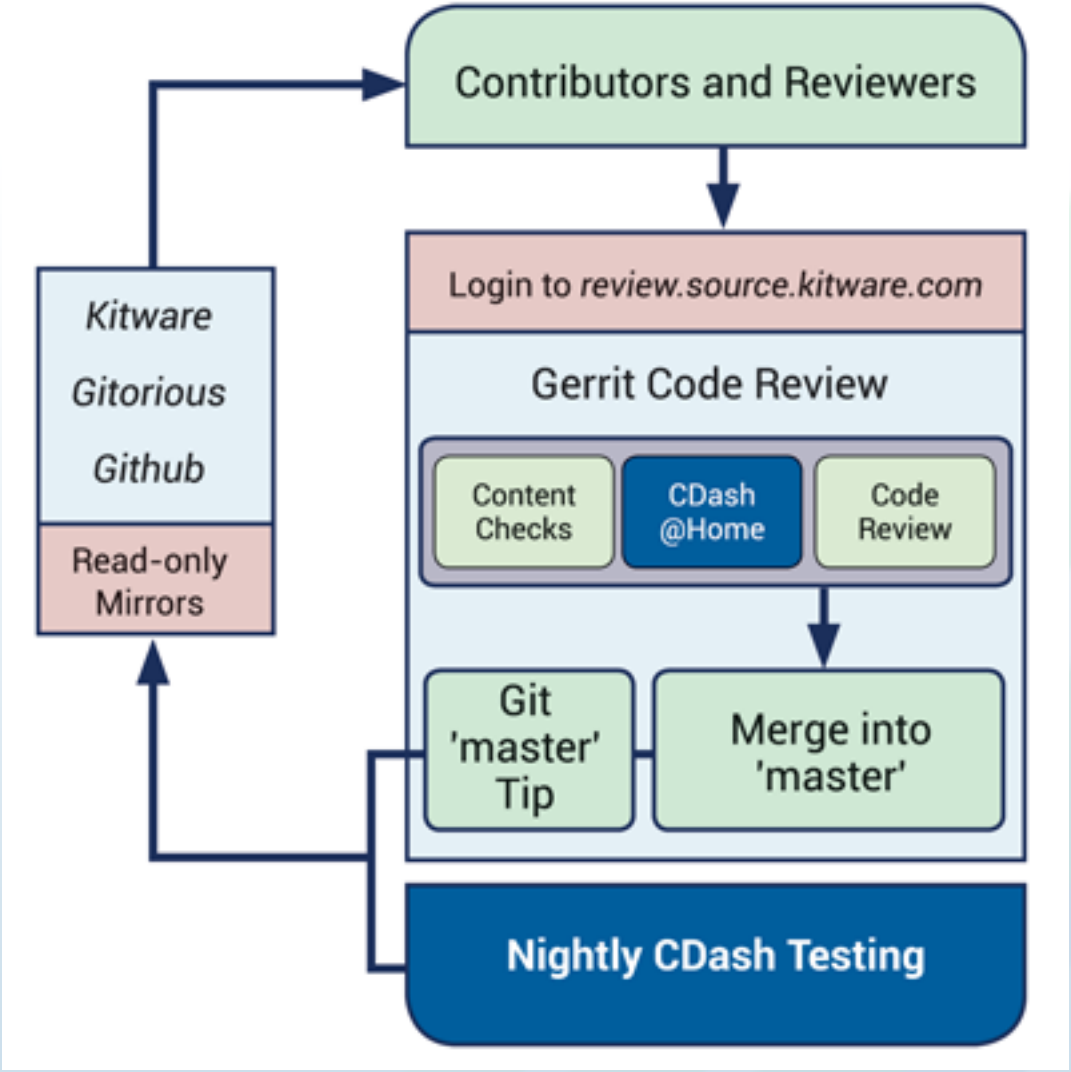
ParaViewWeb and Open Chemistry



Software Process

- Source code publicly hosted using Git
- Gerrit for online 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, and bug tracker

Software Process



Outline

- Introduction
- Kitware
- Open Chemistry
  - Avogadro 2
  - MoleQueue
  - MongoChem
- **The Future**
- Summary

Vision for the Future

- Find partners to develop targeted solutions
- Improved tight integration is needed
  - Computational codes
  - Clusters/supercomputers
  - Data repositories
- Improve and extend client-server architecture
- Co-processing/in-situ visualization/analysis
- Embracing open, semantically rich data
- Address semantic and large data in chemistry

Avogadro: Visualization

- GPGPU accelerated rendering/interop
- More VTK for advanced visualization
- Support for 2D and 3D plots of data
- Optimized data structures
  - Streaming of large data
  - Real-time ray-tracing
- Reworked interface
  - Tighter database/workflow integration

MoleQueue: Complex Jobs

- Tighter integration with remote queues
- Integration with databases
  - Retain full log of computational jobs
  - Trigger actions on completion
- Manage complex jobs
  - Restarts, dependent jobs, triggers
  - Meta-scheduling – choose best resource
  - Classify completed job success/failure/status

MongoChem: Chemistry Data

- Substructure searches
  - Fingerprints support substructure searching
- Tighter integration with applications
  - Communication to search/retrieve/submit
- Easier addition/annotation of data
  - Enable full annotation and searching
- Web frontend wider sharing
- Simple command line tools – batch jobs

Quixote: Parser Technology

- From punch cards and line printers…
- Implement C++ parsers
  - Using regular expressions
  - Provide editor/simulator
  - Easily update parser for new terms
- Dictionaries
  - Documenting the log files
- Facilitating data storage and exchange

Building Community

- Community around chemistry projects
- Using Kitware software process
  - Ensuring quality with continuous testing
  - Code contributions on the web
  - Public mailing lists, bug trackers, code review
- Promoting projects and participation
  - Publication
  - Conferences
  - Workshops
  - Social media

Rethinking Input File Generation

- Can we create a CML representation?
  - Could be loaded directly by some codes
  - Could be translated to input files for others
- Would allow search on input and output
- Could be stored and published
- Make it easier to set up calculations
- Created a more uniform experience
- Input generators currently use JSON rep

Outline

- Introduction
- Kitware
- Open Chemistry
- The Future
- **Summary**

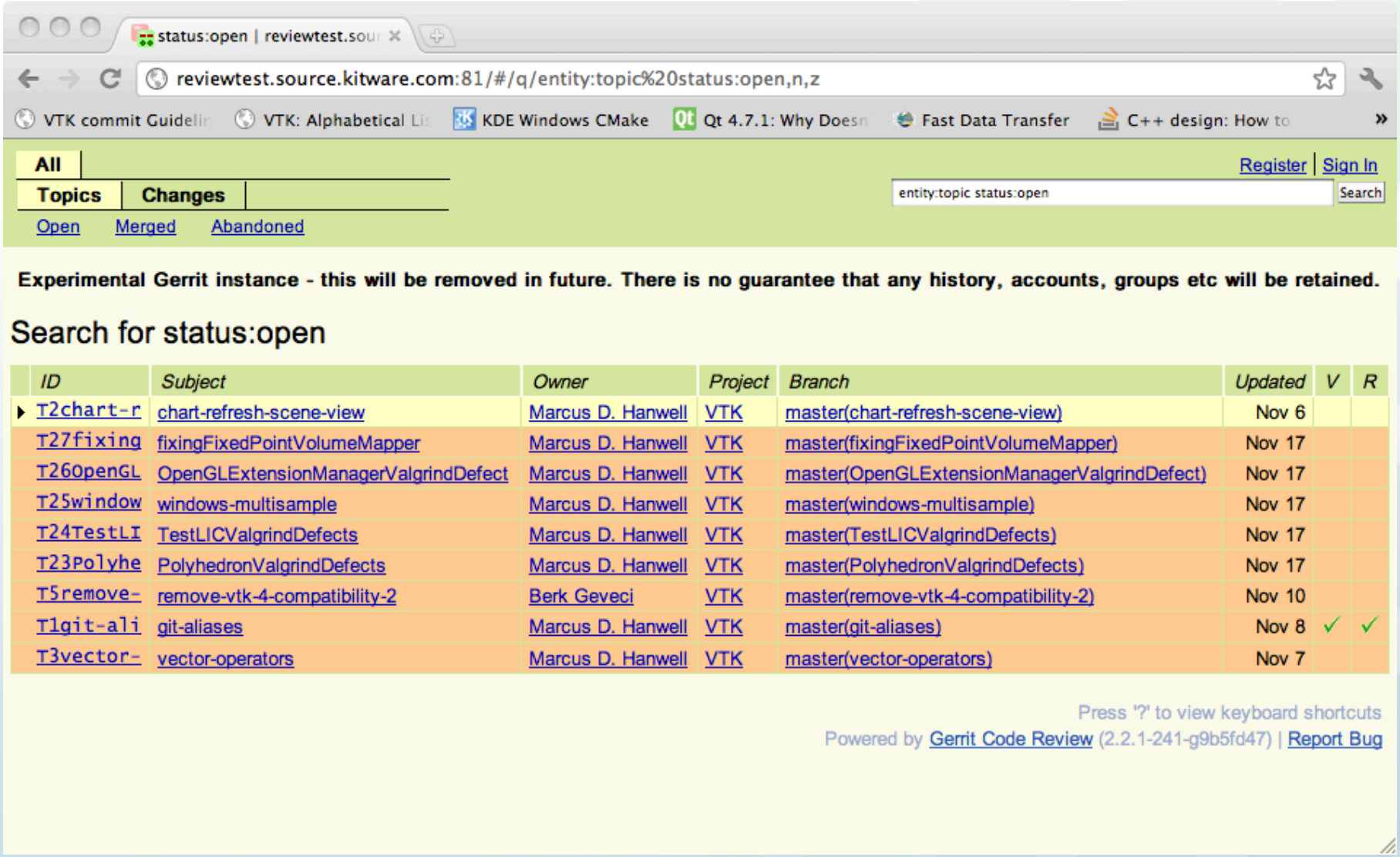
Overview

- Avogadro 2 – molecular editor/analysis
- MoleQueue – manage external code execution
- MongoChem– data management, visualization
- AvogadroLibs – core data structures, algorithms
- VTK – advanced visualization and analysis
- A strong ecosystem for computational chemistry
  - Documentation and training materials
  - Collaborative research and development
  - Work with scientists on real research problems
  - Provision of support and consulting services

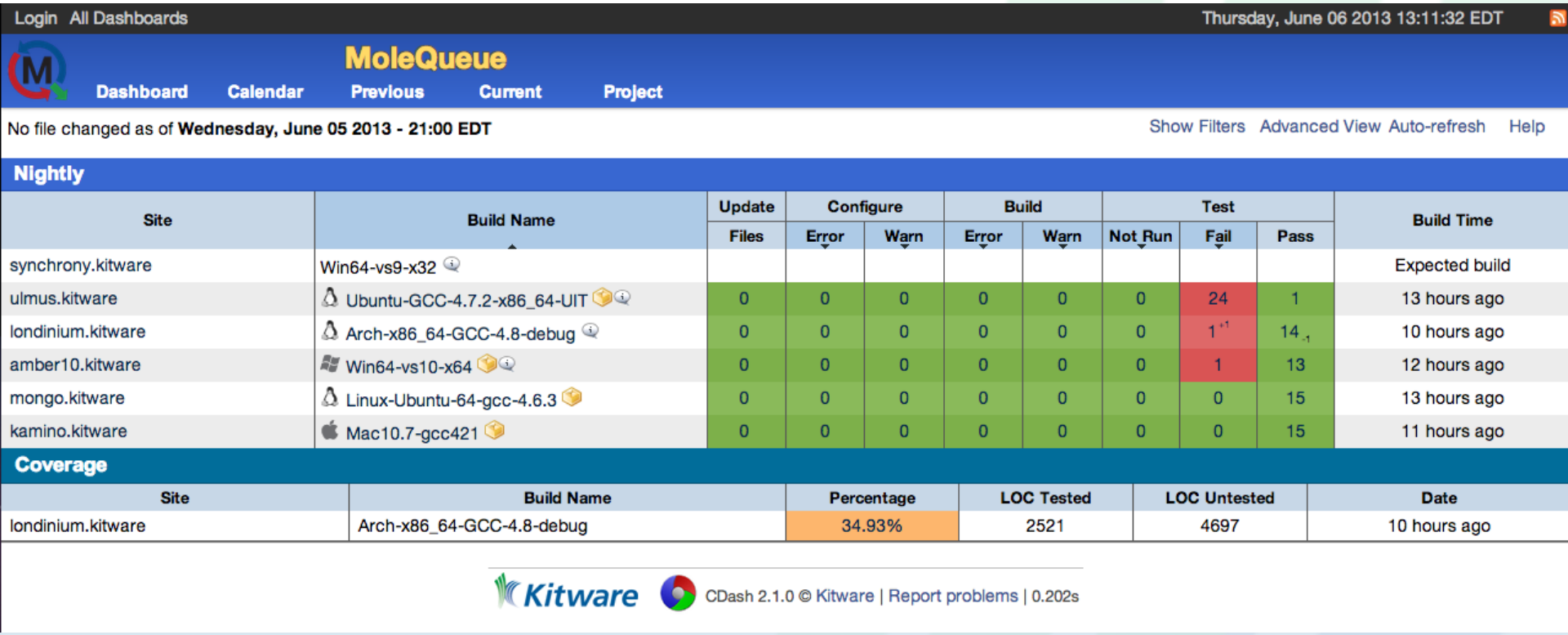
Conclusions

- 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
- Actively seeking collaborators for future work

Gerrit


status:open | reviewtest.sour
reviewtest.source.kitware.com:81/#/q/entity:topic%20status:open,n,z
VTK commit Guidelir
VTK: Alphabetical Lis
KDE Windows CMake
Qt 4.7.1: Why Doesn
Fast Data Transfer
C++ design: How to
All
Topics
Changes
Open
Merged
Abandoned
Register
Sign In
entity:topic status:open
Search
Experimental Gerrit instance - this will be removed in future. There is no guarantee that any history, accounts, groups etc will be retained.
Search for status:open
ID
Subject
Owner
Project
Branch
Updated
V
R
T2chart-r
chart-refresh-scene-view
Marcus D. Hanwell
VTK
master(chart-refresh-scene-view)
Nov 6
T27fixing
fixingFixedPointVolumeMapper
Marcus D. Hanwell
VTK
master(fixingFixedPointVolumeMapper)
Nov 17
T26OpenGL
OpenGLExtensionManagerValgrindDefect
Marcus D. Hanwell
VTK
master(OpenGLExtensionManagerValgrindDefect)
Nov 17
T25window
windows-multisample
Marcus D. Hanwell
VTK
master(windows-multisample)
Nov 17
T24TestLI
TestLICValgrindDefects
Marcus D. Hanwell
VTK
master(TestLICValgrindDefects)
Nov 17
T23Polyhe
PolyhedronValgrindDefects
Marcus D. Hanwell
VTK
master(PolyhedronValgrindDefects)
Nov 17
T5remove-
remove-vtk-4-compatibility-2
Berk Geveci
VTK
master(remove-vtk-4-compatibility-2)
Nov 10
T1git-ali
git-aliases
Marcus D. Hanwell
VTK
master(git-aliases)
Nov 8
T3vector-
vector-operators
Marcus D. Hanwell
VTK
master(vector-operators)
Nov 7
Press '?' to view keyboard shortcuts
Powered by Gerrit Code Review (2.2.1-241-g9b5fd47) | Report Bug

CDash



Login All Dashboards — Thursday, June 06 2013 13:11:32 EDT

MoleQueue

Dashboard Calendar Previous Current Project

No file changed as of **Wednesday, June 05 2013 - 21:00 EDT**

Show Filters Advanced View Auto-refresh Help

Nightly

Site	Build Name	Update	Configure		Build		Test			Build Time
		Files	Error	Warn	Error	Warn	Not Run	Fail	Pass	
synchrony.kitware	Win64-vs9-x32									Expected build
ulmus.kitware	Ubuntu-GCC-4.7.2-x86_64-UIT	0	0	0	0	0	0	24	1	13 hours ago
londinium.kitware	Arch-x86_64-GCC-4.8-debug	0	0	0	0	0	0	1 +1	14 -1	10 hours ago
amber10.kitware	Win64-vs10-x64	0	0	0	0	0	0	1	13	12 hours ago
mongo.kitware	Linux-Ubuntu-64-gcc-4.6.3	0	0	0	0	0	0	0	15	13 hours ago
kamino.kitware	Mac10.7-gcc421	0	0	0	0	0	0	0	15	11 hours ago

Coverage

Site	Build Name	Percentage	LOC Tested	LOC Untested	Date
londinium.kitware	Arch-x86_64-GCC-4.8-debug	34.93%	2521	4697	10 hours ago

Kitware CDash 2.1.0 © Kitware | Report problems | 0.202s

Kitware