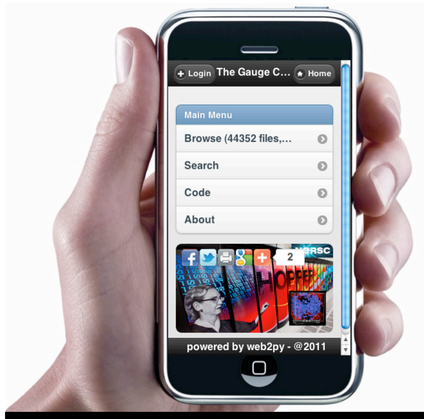
Making Lattice QCD Data Accessible and Organized

James Hetrick (University of the Pacific), David Skinner (LBL), Shreyas Cholia (LBL), Massimo Di Pierro (DePaul University)

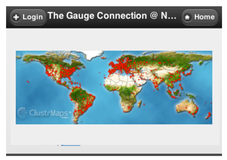
From your phone....

or from your browser



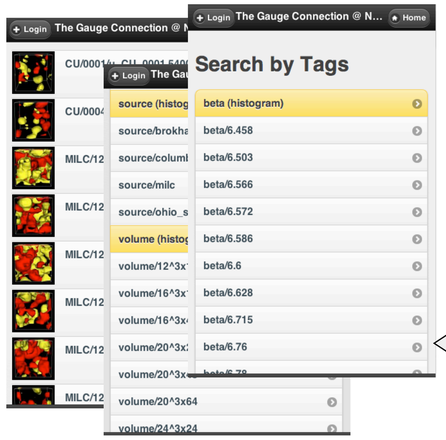
Login

using OpenID



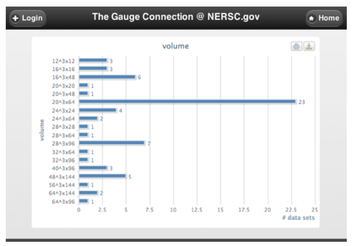
Search

by tag (size, beta, flavor, etc.)



Get Statistics

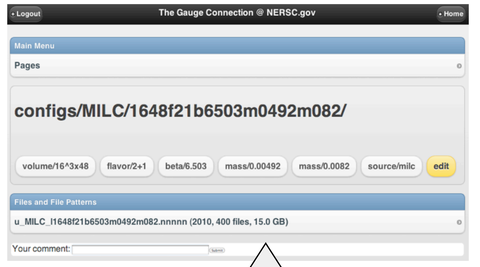
about lattices and about users



data is monitored for changes and tagged automatically based on filename conventions

Browse and Edit

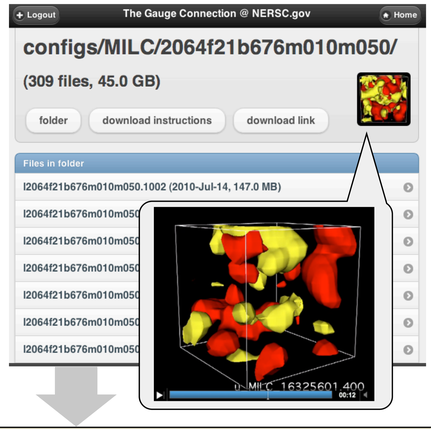
metadata, tags and comment



pages are wikis. add comments including links using wiki syntax and formulas in Latex

Download

and schedule batch downloads/jobs



Abstract and Motivations

The "Gauge Connection" Lattice QCD data archive has been operated by the National Energy Research Scientific Computing (NERSC) center since 1998 and it is now being updated to provide a number of modern features.

With over 16TBytes of data the NERSC archive is one of the largest public repositories of lattice gauge ensembles. Its popularity has largely been due to the simplicity of use of its web interface compared compared with more complex grid based tools.

The features we are adding include the ability search data using tags and to move data easily between the archive and the user's computer, or between two remote computers using both web and grid tools.

The archive is now database driven and its pages are dynamically generated in order to facilitate access to the most recent local and remote data. It also now provides some visulization of the data. New data can easily be uploaded to the FTP server where it is automatically discovered and cataloged.

The archive can now deliver data in a variety of file formats (including ILDG and FermiQCD) by translating data on the fly after dowload.

We are designing the new archive's website to provide more than just data access, including wiki capabilities to annotate the data and link external work derived the from archive data (derived data, software, tutorials, and publications).

Each gauge ensemble is also associated to a discussion forum allowing registered users to add their own comments.

The archive has a more sophisticated access control mechanism and users can have four possible roles: administrators (can manage every aspect), editors (can edit the wiki pages linked to the data), registered users (can download data and comment), anonymous visitors (can browse the wiki pages and query the data by tags).

Users can login using their existing OpenID credentials: for example using their Google account, in addition to grid credentials.

Basic Features

The NERSC data is stored on the High Performance Storage System (HPSS), a tape storage system developed by a collaboration between five Department of Energy laboratories and IBM with significant contribution form Universities.

HPSS stores more than 1Peta Byte of data, including 16 Tera Bytes of Lattice QCD data. HPSS has an FTP interface, a GridFTP interface, a web interface and a Globus Online Interface.

The new web interface is designed to mimic the iPhone interface and works on both regular computers and mobile devices.

The system performs nightly introspection of the FTP folder structure, reproduces this folder structure in the database, automatically tags data by parsing the file names, groups files with similar name patterns, and publishes the data online.

For each folder on the FTP server a corresponding web page is generated dynamically. The web structure has the same hierarchy as the folder structure on the server. Every page is editable using a wiki syntax similar to wikipedia and registered users can comment on the pages. Both the wiki and comments allow Latex syntax for formulas.

Gauge ensembles can be searched by direct browsing of the folder structure or by tag (contributing collaboration, lattice size, beta value, dynamical quark masses, etc.).

The system generates interactive charts with statistics by tags.

Some ensembles have been processed off-line to generate meta-data such as a topological charge densities and have been linked to images and movies of said topological charge density.

The system is data agnostic and in principle it can store more than gauge ensembles. It can store, for example, quark propagators and eigenvalues, examples of which are already in the system.

The system uses standard third party web analytics tools to track usage and geo-tag visitors on a map.

System Design

The system is based on web2py, a framework for rapid development of secure database driven web applications. It is written in Python and supports many standard databases including Sqlite, MySQL, PostgreSQL, Oracle, MSSQL, informix, DB2, Sybase, Firebird, and Google Bit Table. The SQL is generated automatically.

The system uses jQuery Mobile for page layout, Google Chart API for Latex rendering and matplotlib for plots and charts.

The visualizations of topological charge density are produced offline using Visit (LLNL) and the FermiQCD Visualization Toolkit.

The system has a module Model-View-Controller design which separates the data representation from the data presentation and from the application workflow. This makes the code compact, easy to maintain and modify. It includes a web based IDE, a web based database management tool and internationalization capabilities.

The system itself is not domain specific and has no knowledge about QCD files and conventions. The domain specific knowledge is in a separate script that runs in background and populates the database from the file/folder structure. This means the system can be used to publish data of a different nature with minimal work.

The system retains the original ability to download individual files using the web interface. Since these files can be big, to prevent Denial of Service attacks, the download requires login.

The system exposes RESTful web services APIs based on the JSON protocol which can be used to access the data programmatically.

Some gauge ensembles are comprised of thousands of files. To download them in batch we provide a script written in Python that can query the server for all fields in an ensemble, downloads them one by one, converts them to the required target format and precision, and logs the completed work. This allows arbitrary resume capabilities and avoids un-necessary duplication of work.

```
$ qcdutils.py http://qcd.nersc.gov/nersc/api/files/demo
http://qcd.nersc.org/nersc/api/files/demo
target folder: demo
total files to download: 1
downloading demo.nersc
demo.nersc 100% |###############| Time: 00:00:00 654.52 K/s
completed download: 1/1

$ qcdutils.py --covert ildg --float demo/demo.nersc
converting: demo/demo.nersc -> demo/demo.nersc.ildg
  (precision: f, size: 4x8x8x8)
100% |##################################|

$ qcdutils.py demo/qcdutils.catalog.db
demo.nersc created on 2011-06-17T13:42:30.876812
   [14e7cf9106bfcc16388aeac285ccdad9]
demo.nersc.ildg created on 2011-06-17T13:42:34.424604
   [5a1ae13ddd5cab7ddfe1b17454822ff5]
```

Outlook and Acknowledgments

The system allows the user to register any URL and dynamically generates buttons that, when pressed, pass a link to the data to the associated URL. This will allows the creation of third party web services that can be feed data directly from the new NERSC web interface allowing for decentralization of services. We can provide tools to help create such services that interface on the one side to the NERSC archive and on the other side, to your PBS job submission queue.

We envision a future when the different research groups will provide their computing capabilities and their lattice QCD algorithms as web services for the consumption of other members of the collaboration. The NERSC site provides more than just data for these collaboration but also an infrastructure to register those third party services in a transparent way.

Acknowledgements

This work was funded by Department of Energy grant DEFC02-06ER41441 and by National Science Foundation grant 0970137.