Data Handling Summary

- A condensed (and updated) version of talk from November:
  - https://confluence.slac.stanford.edu/x/tgvKBg
    - No pictures, no background info, small fonts
- Status and Plans
  - People, Goals and Techniques
  - Major Experiments
    - Fermi
    - EXO
    - LSST Camera Control System
    - ILC/LCIO/lcsim
  - Cross Experiment Projects
  - Service Level Catalog, Support Levels
  - Future Innovation
  - Decision points
  - Manpower roadmap

People, Goals, Techniqes

- People
  - Karen Heidenreich, Chalotte Hee, Tony Johnson, Dimitry Onoprienko, Max Turri, Brian Van Klaveren
- Goals
  - Long term:
    - Well supported key software infrastructure components ready for adoption by experiments at SLAC and other universities and laboratories
  - Strategic:
    - Invest in new technologies for future experiments
    - Continue to encourage software reuse between experiments at SLAC
    - Build collaborations with the Computing Department, LCLS and other laboratories
  - Tactical
    - Support existing/planned SLAC experiments
- Techniques
  - Core Strengths:
    - Wep Application Development, Data Quality Monitoring, Analysis and Visualization (Web+GUI), Data Access, Experiment and User Support, Developer/Collaborative tools and techniques
  - Design Methodology
    - Design for experiment independence/reuse
    - Small modular interoperable components with well defined interfaces
    - Adopt industry/community standards whereever possible

Fermi Large Area Telescope

Major Data Handling group contributions

- **Data handling coordinator (Tony)**
- **Web Applications**
  - Portal, Login, Group Manager, Shift System, Resource Monitoring, Data Processing, Speakers Bureau, Run Quality, System Tests, …
  - Data Quality Monitoring/Trending (time histories)
  - Automated Science Processing
- **Web Infrastructure**
- **Automated workflow engine (pipeline)**
- **Data Storage and Access**
  - Use of xrootd (with Andy and Wilko)
  - Data Catalog + Download Manager
  - Astro Server and Skimmer (event filtering)
  - WIRED Event Display

Ongoing/recent/planned work

- **Maintenance of existing projects**
  - Upgrade maven infrastructure (Max, maven 3 ongoing)
  - Upgrade tomcat infrastructure (Charlotte, tomcat 7, IIS ongoing)
  - Support/Upgrade Java/tools installation at SLAC (Charlotte, java 7, ongoing)
  - Oracle upgrades (Oracle 11 ongoing)
  - Update Fermi projects to use same base infrastructure as other experiments (Max, Charlotte, Karen, done)
  - Astro-server data reloads (Brian, pass 7, 8, ongoing)
  - WIRED event display longstanding bug fixes (Dimitry, done)
  - User management (Karen)
- **Active/Pending projects**
  - Pipeline (Brian)
    - Extension to Grid, Mixed Mode
    - Web performance improvements
  - Speakers Bureau improvement/extensions (Karen)
  - Integration with campus web tools (Karen)
  - Integrate CAS and Crowd logins (Max)
  - Improve trending performance (Brian, Max)
  - Group/User manager upgrade (Karen, Charlotte, Max)
  - Nagios Migration to SCCS (Charlotte)
  - Web interface improvements
    - Dat a catalog (Brian)
    - Astro server/Extended analysis (?)

EXO

Major Data Handling Group Contributions

- **Computing Coordinator** (Tony)
- **DAQ** (web based GUI based on GWT)
  - With JJ and TonyW
- **Online data quality monitoring** (AIDA, Remote AIDA, Web Infrastructure) and **Online Event Display** (Tony, Max, New Plotter)
- **Data Transport**, Data (re)Processing, MC Infrastructure (Pipeline), disk/tape at SLAC (xrootd)
- **Web Applications** for
  - Trending, Data Quality, Shifts, Portal, User/Group Manager, Login, Data Mover, Data Catalog, MC Infrastructure, Monitoring ...
  - Extensive reuse of Fermi/SRS/FreeHEP tools (Max, Karen, Charlotte)
- **Offline** (Tony)
  - Build system (make)
  - Subversion (sventon)/Automated build (hudson)
  - Documentation (confluence), JIRA
  - Conditions system (with Joanne Bogart)
- **Security, Firewall at WIPP** (Tony)
  - With Matthias and TonyW

Ongoing/recent/planned work

- Installation of new computers at WIPP (TonyJ, TonyW, Matthias - coordination)
  - Reliable, Secure operation thru 2016
- Improve effectiveness of online/offline monitoring (mostly new plots, small improvements to infrastructure)
  - Automatic checks
- MC Infrastructure (done)
- Slowcontrols/Web de-duplication (Matthias)
- User management (Karen)
- Final upgrade of main data processing pipeline
  - Switch to compressed data format
  - Expanded use of xrootd
  - Document so others can take over
- Web interface for conditions system (Karen, Charlotte)
- Disk handling, processing, reprocessing (Tony)
- Maintenance (see Fermi)
- Computing coordination
  - Disk/tape purchases.
  - Integration with new SLAC cluster
  - 6 hours/week meetings
  - Being woken up at 6am because someone powered off the data quality computer @WIPP

LSST

Major Data Handling Group Contributions

- Led by Max
- with Tofigh, Stuart, Owen and Paris Group

Camera Control System

- Understand existing infrastructure
  - Work with Paris group on redesign, bug fixes, improvements
- Move to supportable build/management/test system
- Developed GUI's for shutter/carousel (Tofigh, Tony)
- Demoed working system interfaced to prototype hardware (Tofigh, Owen)
- Engineering console
  - based on JAS plugin framework
- Developing web interface/trending system
- Beginning to support first sets of users
  - Established user support mailing list
- Developer/User workshops
  - SLAC in October 2011
    - Developed roadmap for next 6 months
  - Brookhaven in March 2012

Longer term involvement in LSST?

- System engineering for observatory
- Web interfaces for science

Ongoing/recent/planned work

- Significant development project
  - Expect to take significant manpower for >3 years
- Weekly EVO developer meetings
  - Debug/iterate on core infrastructure
- Shutter controls/GUI (new plotter)
  - Interfaced to real hardware
- Development of refrigerator test stand
  - To be used Feb/Mar 2012
  - Console with integrated trending, custom GUI, Web Server, restful interfaces, AIDA
- Filter Changer test stand (Summer)
- CCD test stand (?)
- DAQ interface (?)
- CCS/Observatory interface (?)

ILC/LCIO/lcsim/…

- (See Norman's, Homer's talk)
- LCIO = joint project with DESY to define standard IO format for ILC detector studies
  - Provided leadership on development of this library
    - Norman, Tony on ILC common software task force
  - Uses subversion, hudson, JIRA, confluence
- lcsim = reconstruction and analysis framework
  - Lightweight, extensible, easy-to-learn framework
    - Developed for ~15 years with contributions from ~100 people
    - Shares much of the same "data handling" infrastructure we use for other experiments
  - Specific requests for support
    - Add support for LCIO random access (for SiD)
    - Database interface for lcsim conditions system (for HPS)

Cross Experiment Tools

Web Infrastructure/Data Quality/Tools

- Common web infrastructure
  - Based on tomcat (J2EE) web servers
  - Common authentication/authorization scheme
    - Supports SLAC windows or unix ids (via CAS/kerberos)
      - May extend to support crowd to integrate confluence ids
      - Close integration with Crowd/Group Manager
- Many cross-experiment web applications
  - User database (Fermi?, EXO)
    - Meta-data customizable by project, can track history of collaborators, control mailing lists, generate publication lists
  - Shift calendar (Fermi, EXO)
  - Data Quality, Trending (time histories)
    - Toolkit, interfaced to oracle, mysql, root, fits
    - Can develop project/experiment specific applications
      - 50+ apps for EXO/Fermi, including user developed apps

Cross Experiment Tools
Pipeline + Data Access

- Pipeline (data flow engine)
  - Allows complex, parallel workflows to be defined
  - Data-processing history in database
  - Allows jobs to be ”rolled back” manually or automatically
  - Extensive web interface, plus command line interface
- Data catalog (data access)
  - Heirachical database for keeeping track of data
    - Storage independent, particularly useful with xrootd
    - ”Crawler” which verifies data integrity
    - Allows arbitrary meta-data to be associated with datasets
      - At registration time, or extracted by crawler
  - Web interface for browsing/selecting/fetching data
    - Download manager, Skimmers, Event Displays
  - Command line interface
- Ongoing/recent/planned work
  - Fermi dependencies removed
    - Pipeline+Data Catalog used by EXO (MC, Data Processing), CDMS (MC), CTA (SLAC MC)
      - EXO probably has more pipeline users that Fermi
        - Recently reprocessing data each night
    - Pipeline extended to work with
      - Grid Engine (IN2P3)
      - Condor (SMU)
      - Work ongoing to make ”generic” Grid interface (EGEE/Dirac)
  - Possible improvements
    - Current web interface ”low tech”
      - Pipeline web interface performance for Fermi poor
        - Web interface could be improved
          - Dynamic loading (HTML5)
          - Simplification
      - Add json interfaces
        - Better integration with python etc
      - Better web based search
    - Remove Oracle dependence
    - Better documentation
  - Should we be encouraging other users?
    - Other SLAC experiments
    - Interest expressed by others James Webb Space Telescope, CTA
    - How to evaluate how much effort it is worth?

Cross Experiment Tools
Developer/Collaborative Tools

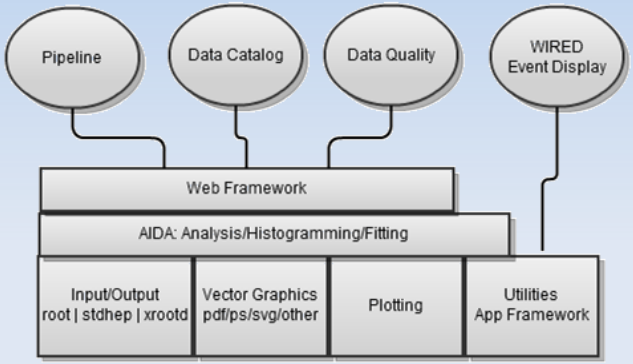
- Maven
  - Java based project management/build system (Fermi, EXO, LSST, lcsim)
- Subversion
  - svn.slac.stanford.edu, svn.freehep.org
  - Web Interface to subversion (sventon)
  - E-mail notification on checkin (svnspam)
  - Integrated crowd login (in progress)
  - (Exo, Fermi, CCS, theory, CTA)
- CVS
  - cvsspam, cvs.freehep.org (lcsim, hps, etc), web interface (currently dead)
- Hudson
  - Automated build system (EXO, CCS, Fermi, lcsim, CTA)
- Whole is greater than sum of parts
  - IDE integration
  - All of these systems interoperate
- JIRA/Confluence/forum
  - Support now handed to computing division
  - But still help with user requests, configuration, migration to newer versions
    - e.g. LaTex -> MathJAX
    - E-mail support for forum.slac.stanford.edu?
- Atlassian Fisheye
  - Works for CVS, Subversion, GIT, ….
  - Would replace sventon+svnspam+cvsspam
  - Crowd integration built in
  - **Need to decide if/how to support this**
    - **Ideal would be collaboration with computing division c.f. confluence**

Product/Service support levels

- Need to clarify product/service lifecycle
  - Create and maintain service catalog
    - Needs to take into account coupling of projects
    - Needs to support evolution/revolution
  - Advertise it (to attract new customers like HPS, CDMS, …)
- Possible support levels
  - Archived
    - Projects/services which are no longer in use or supported. Kept for archival purposes in case code is useful for future products or other users. No support/updates or bug fixes.
  - Legacy
    - Product now supported only for use by specific experiments, possibly for a limited time period. Where the time period is limited some recommended transition path to newer products may be provided. In exceptional cases products could move from Limited to Active, but the more usual path would be to move from Limited to Archived. Only critical/emergency bug fixes applied. Adoption by new experiments encouraged only after specific discussion of requirements/support level/timeline.
  - Limited
    - Product developed for a specific experiment. Release schedule as agreed with experiment.
  - Active
    - Actively supported for use by experiments at SLAC. SLAC affiliated experiments actively encouraged to adopt product/service. Regular feature and bug fix releases.
  - Product
    - As Active but supported for use at SLAC or beyond.

New Initiatives

- Success of supporting multiple experiments depends on reuse of components



  - Some of these components are very old
  - Need to renew and move forward while maintaining support
    - Need 5+ year timeline
    - Unplanned growth leads to wasted time/effort
  - To be effective new initiatives need at least 2-3 people working together

- Directions:
  - New Plotter
    - Technology review, Pilot projects
    - HTML5 Canvas, GWT, Google visualization API
  - Interactive web applications
    - HTML5, Cross device
    - “cloud” based access to data
    - Collaborative
    - Virtual Observatory
  - Need to find small demonstration projects
    - E.g. resource manager
  - Need to identify potential partners/users
    - LSST, Fermi…

Decision Points

- Decide on support levels for existing products/services
- Pipeline
  - Decide at what level to encourage new users (at SLAC, beyond)
- Data catalog plans
  - Generalize web front-end for use with other products?
- Fisheye
  - Need to decide whether to move forward and whether to partner with CD apps group
- Future support for lcsim
- Manpower/funding for renewal/new initiatives/collaborations

Manpower Profile

- EXO decreasing starting in 3 to 6 months (but expect the unexpected). Continued maintenance required.
- Fermi maintenance stable (trending, pipeline, astro server, data catalog, campus migration). WIRED completed.
- LSST increasing (additional manpower coming?)
- Pipeline/Data Catalog/Web Tools – unknown
  - Web apps/user support – some room for expansion
- lcsim – small but unknown
- Innovation/Planning/Core infrastructure renewal – need critical mass to make this happen