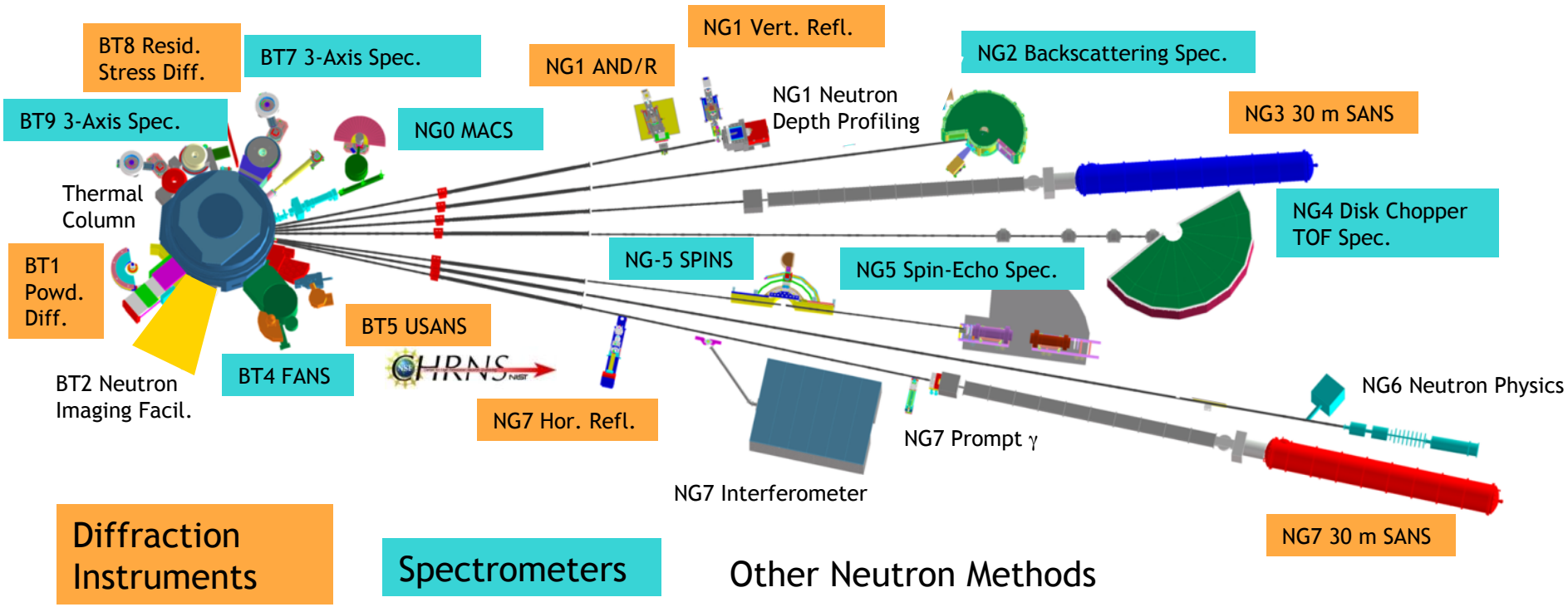
NCNR Raw Data Archive System

Nicholas C. Maliszewskyj
NIST Center for Neutron Research
Research Facility Operations Group

- Facility Overview
- Data Management Considerations
- Evolution of the archive
- Archive Construction and Operation
- Lessons learned

NIST Center for Neutron Research

NIST Center for Neutron Research



Diffraction Instruments

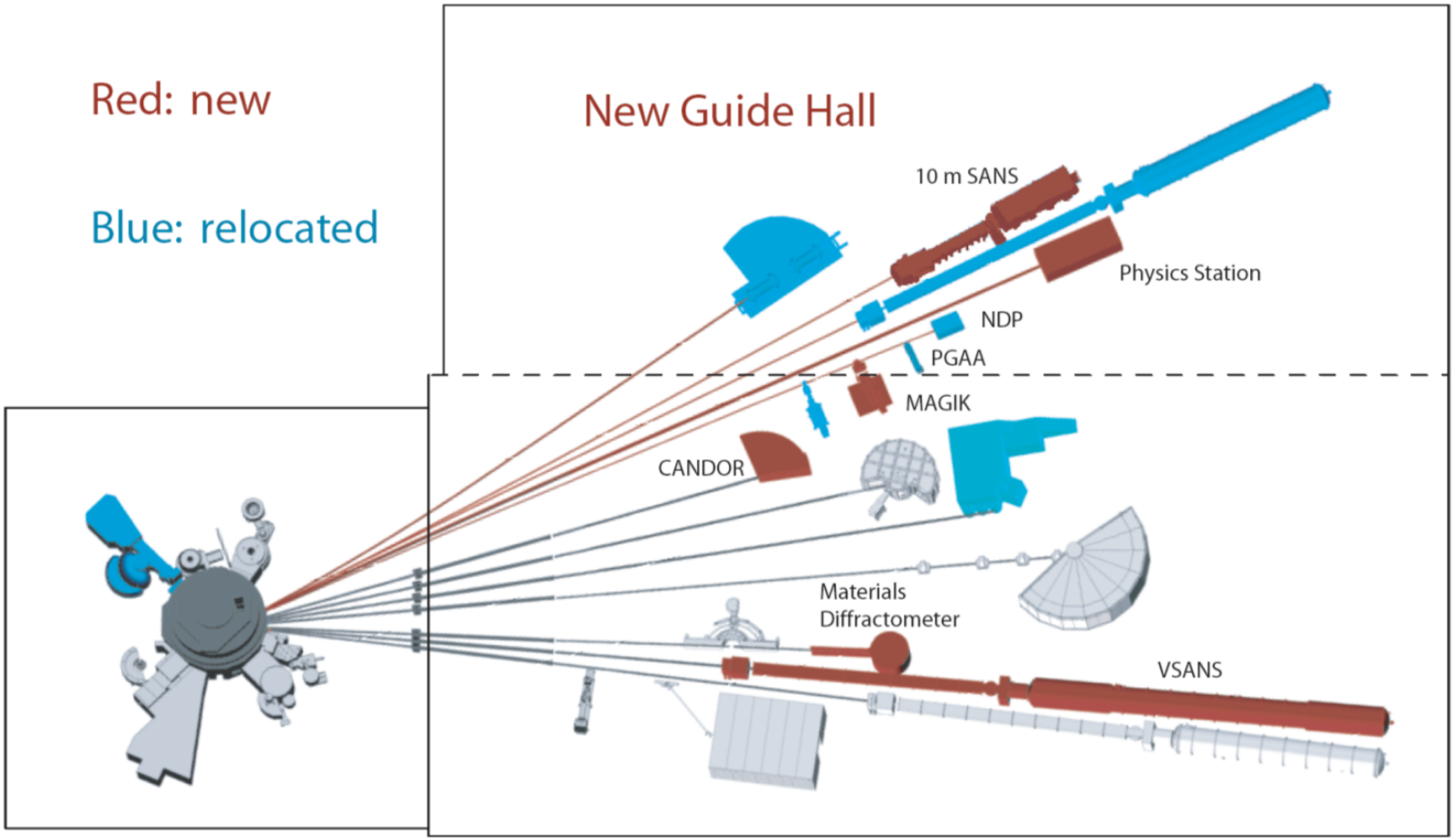
Spectrometers

Other Neutron Methods

20 MW Reactor in service since 1967
21 instruments

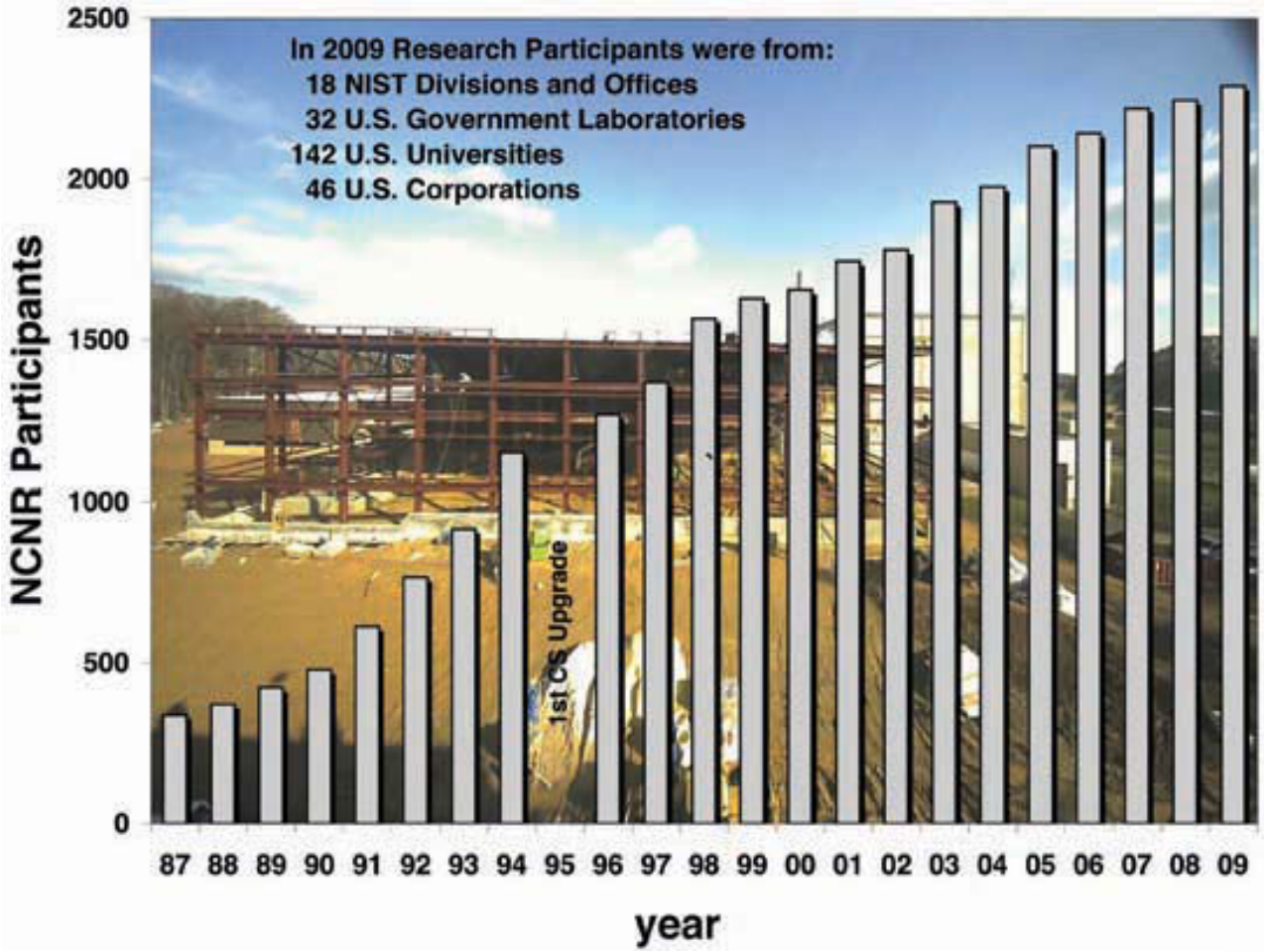
NIST Center for Neutron Research

NCNR Expansion



NIST Center for Neutron Research

NCNR User Program



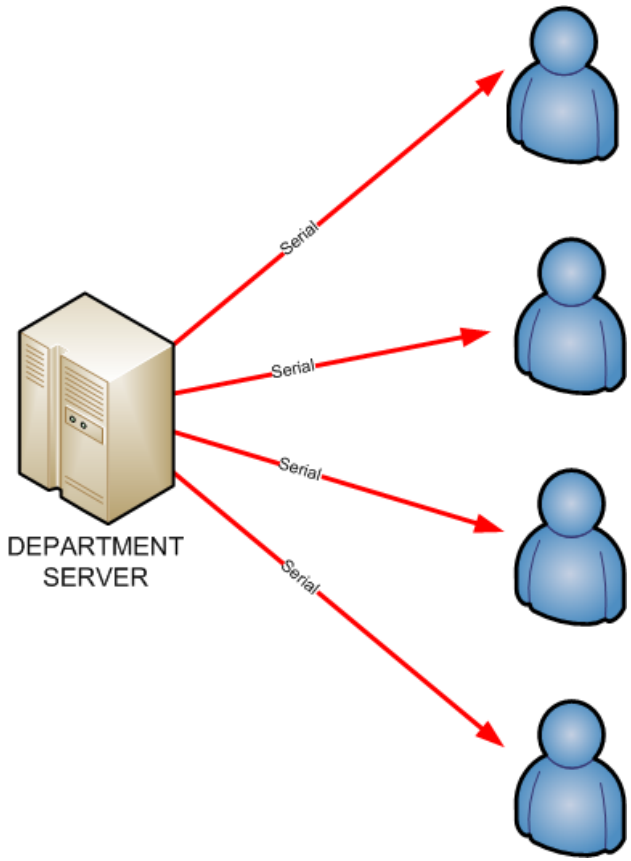
NIST Center for Neutron Research

Data Management Considerations

- Policy
- Culture
- Technology
- Resources

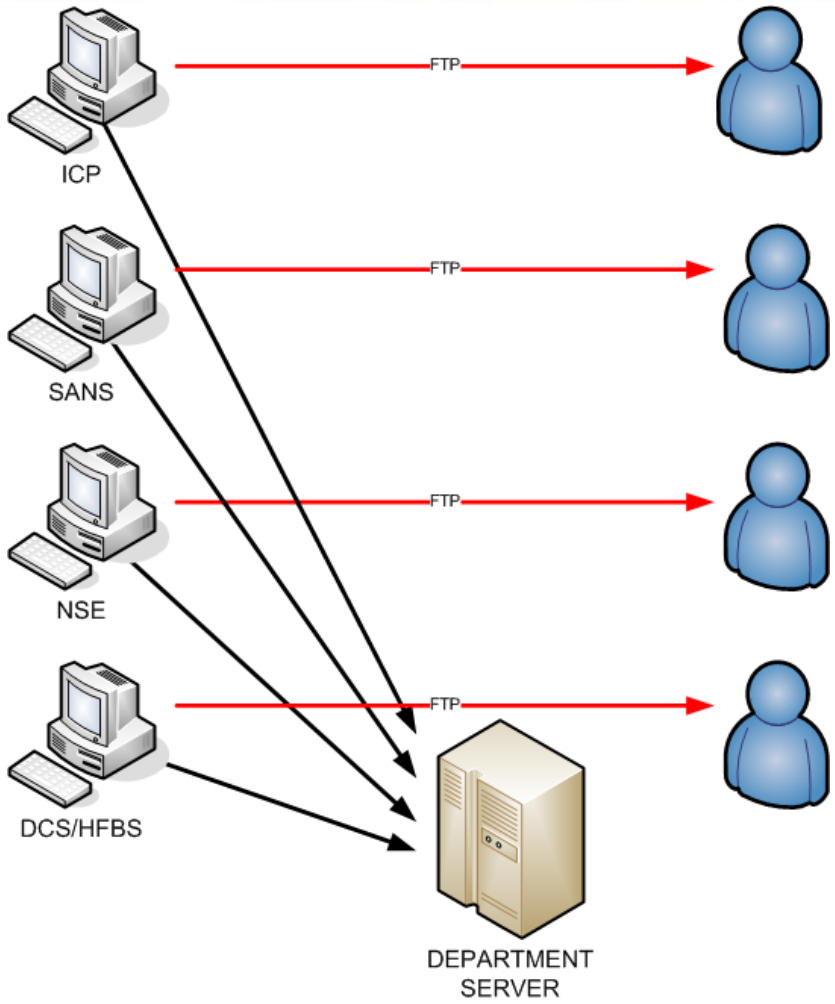
NIST Center for Neutron Research

The Early Years (1970s – 1980s)



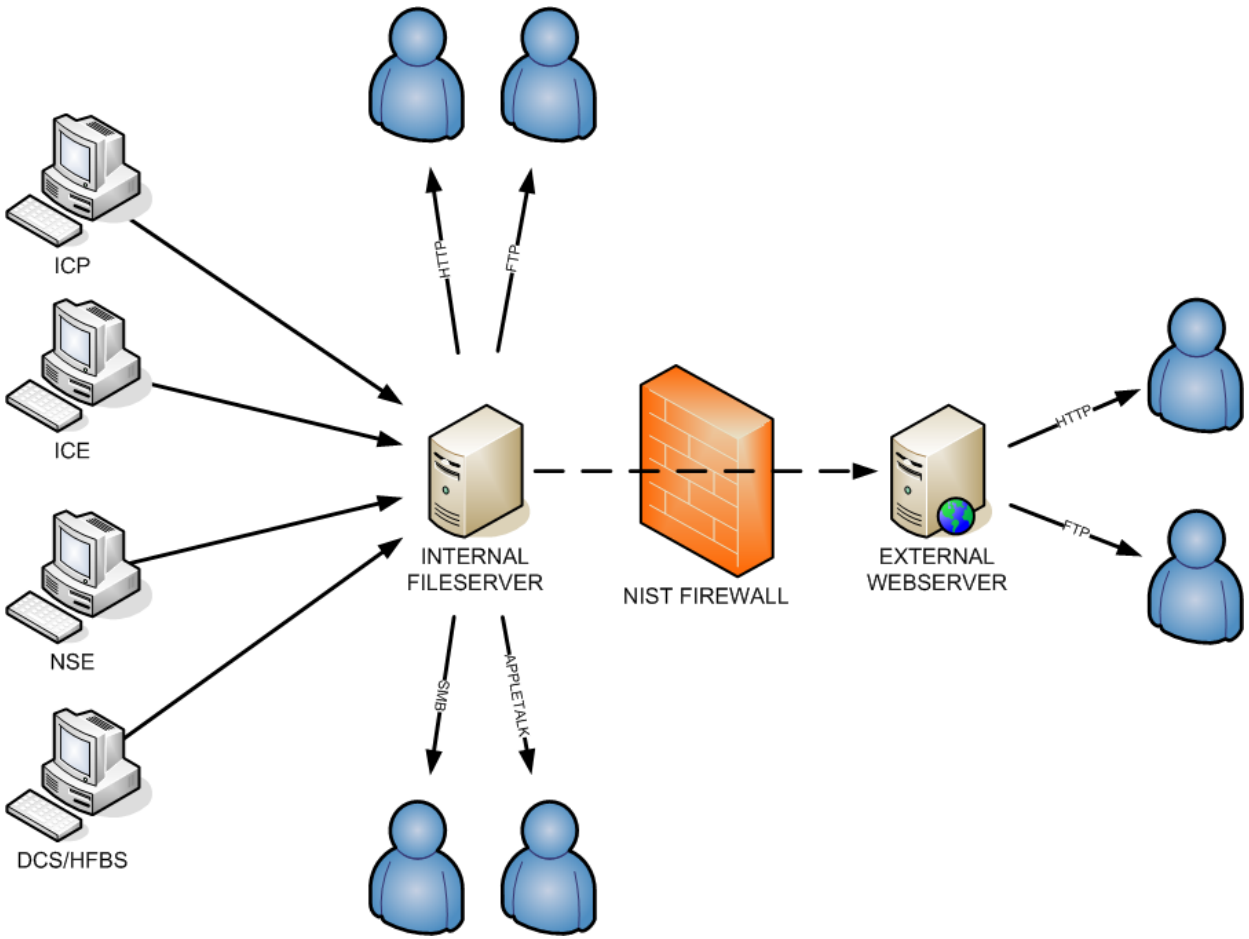
NIST Center for Neutron Research

The Recent Past (1990)



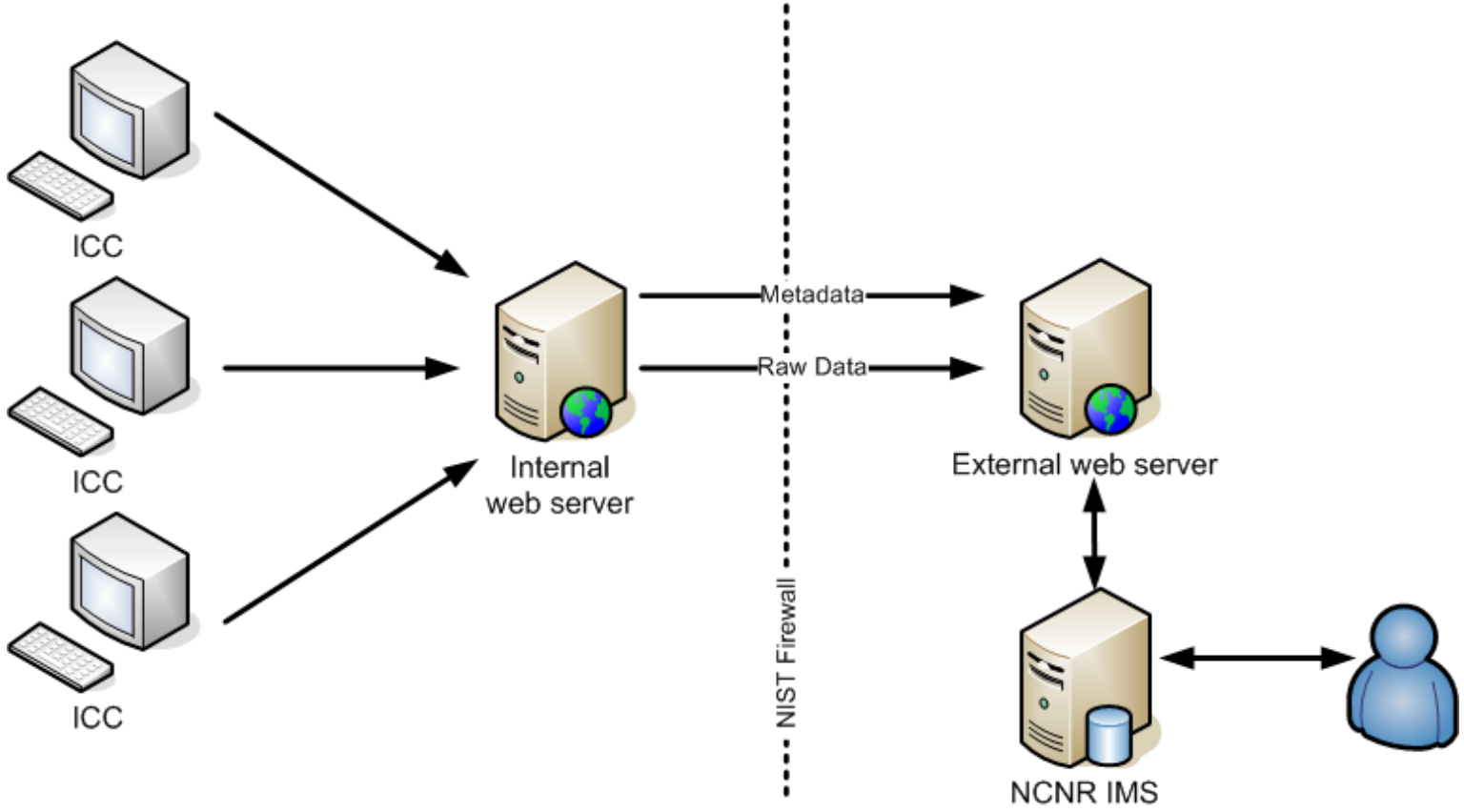
NIST Center for Neutron Research

The Present (2010)



NIST Center for Neutron Research

The Vision



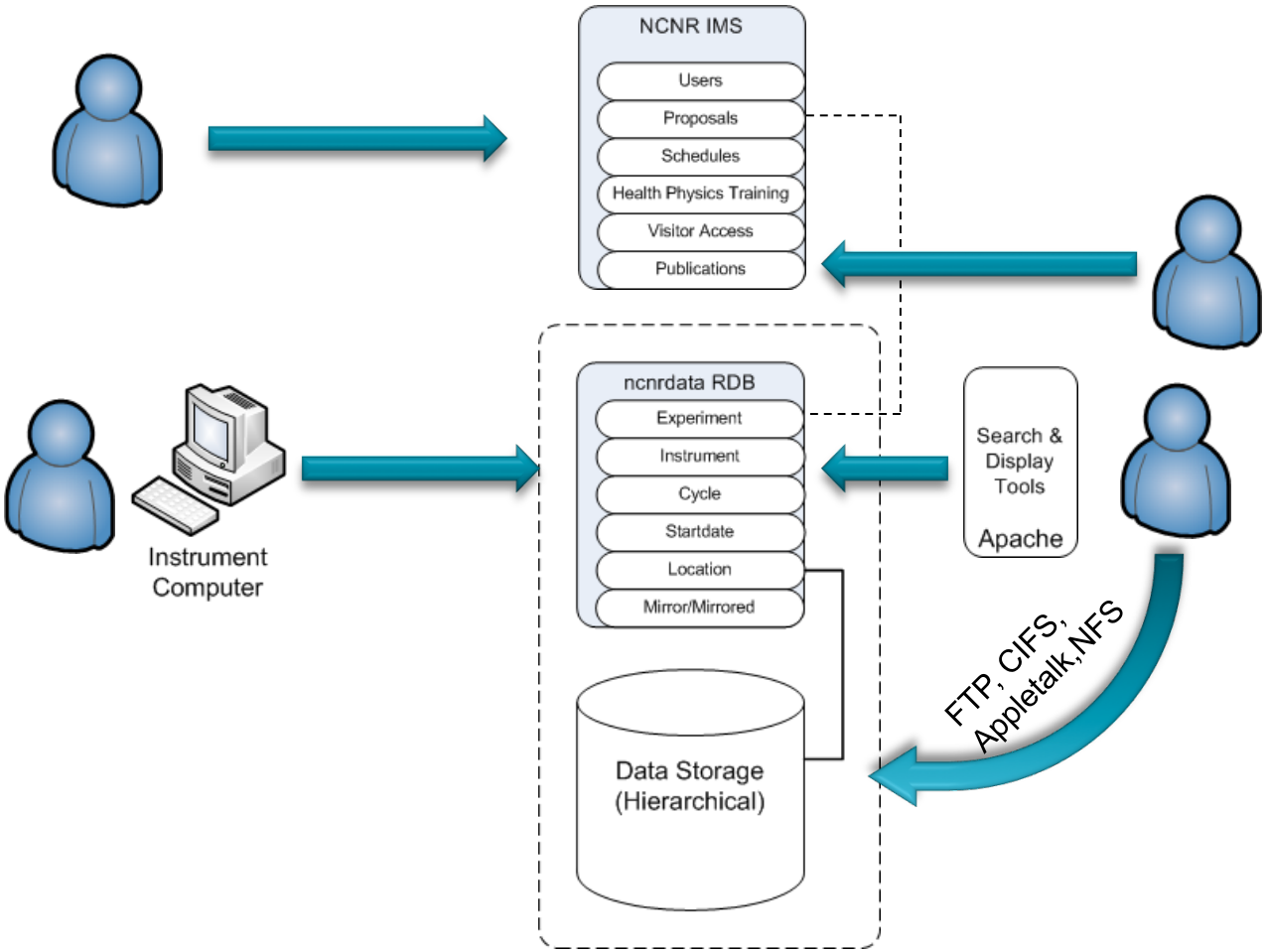
NIST Center for Neutron Research

Tools

- Linux (Fedora variants) – Platform
- MySQL – Metadata RDBMS
- Perl/Net::SFTP, DBI, DBD/MySQL – Scripts
- PHP – Search pages
- Apache, VsFTP, Samba, Appletalk – Content delivery

NIST Center for Neutron Research

Experiment Lifecycle



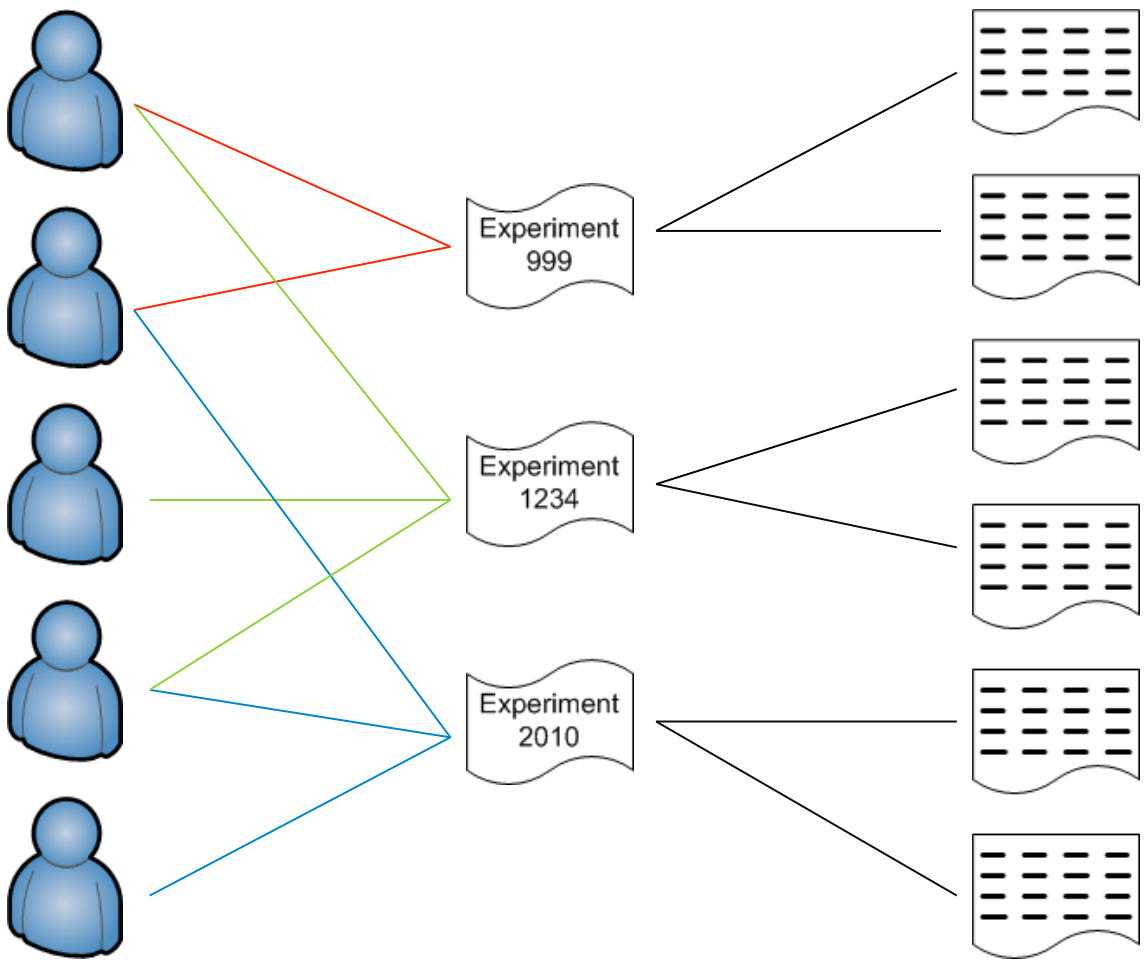
NIST Center for Neutron Research

Archive process

- Instrument Control Software (all suites)
  - Writes a manifest of data files written
- Archive Script (per instrument control suite)
  - Retrieves manifest file from instrument computer
  - Retrieves all files on the manifest; skipping those already present in repository (or which haven't changed)
    - Files organized in repository by instrument, reactor cycle, and working directory
  - Harvests metadata from manifest and raw data
  - Assigns experiment number based on instrument schedule
  - Writes metadata to database
- Mirror Script
  - Periodically copies raw data to external fileserver and mirrors metadata to external database

NIST Center for Neutron Research

Establishing File Ownership



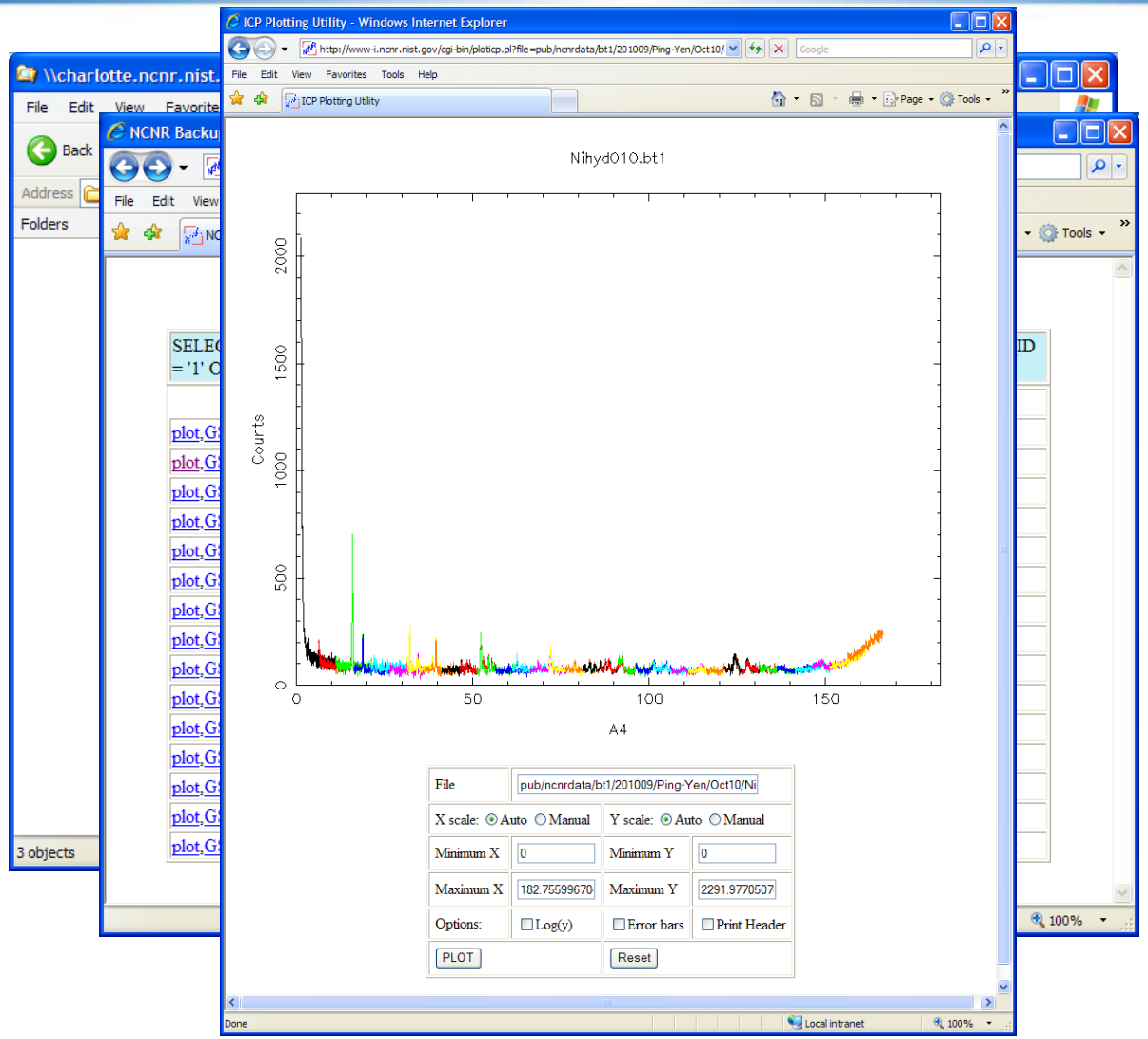
NIST Center for Neutron Research

Exporting data to public

- No access controls
  - Determined to be useful, but potentially a logistical nightmare for the benefit obtained
- A mirror script checks database for files marked to be mirrored
  - Those files and associated metadata are written to a server in public webspace
  - Those files are marked as having been mirrored in database
- By default, all datasets are marked to be mirrored.
  - NCNR Policy announced August 2010
  - Obama Administration Open Government Directive (Dec 2009)
- Users may opt out of the mirror process for a period of up to 18 months

NIST Center for Neutron Research

Accessing the repository



- File exploration tools
- Searchable catalog
- Web based tools

NIST Center for Neutron Research

Capturing Metadata

- In theory –
  - End users provide context for the measurement before the measurement starts
  - Experiment metadata can be harvested from other systems (proposal/ scheduling)
  - The published schedule is authoritative information
- In practice –
  - Most end users skip past dialogs requesting metadata and instead keep their own notes or generate comments meaningful to them
  - Metadata required by the proposal process is the most complete, but you may not be able to fully interoperate with that or other systems
  - The as-run schedule can differ from the published schedule by hours or days

NIST Center for Neutron Research

Delivering content

- In theory
  - Access controls are important and prevent unauthorized access to data
  - Serving raw data from central location is easier
  - Archived raw data are immutable

- In practice
  - Convenience trumps access controls
  - Many old hands don't adapt easily to new schemes for sharing
  - If data served centrally, archival copies are expected to be "hot off the presses"

NIST Center for Neutron Research

Role of Data Format

- In Theory
  - Self describing, extensible formats will encapsulate all necessary information required to reduce and analyze the raw data
  - The entire instrument suite at a facility should write the same data format
- In Practice
  - That information is only as good as our means of obtaining it
  - Changing data formats on “production” instruments requires a great deal of coordination between producers and consumers of data

NIST Center for Neutron Research

What's next

- Better integration with other business systems
  - NCNR User Portal?
- Unified data format
  - Foothold established following 2011 outage?
- Whatever is needed as it is needed

NIST Center for Neutron Research