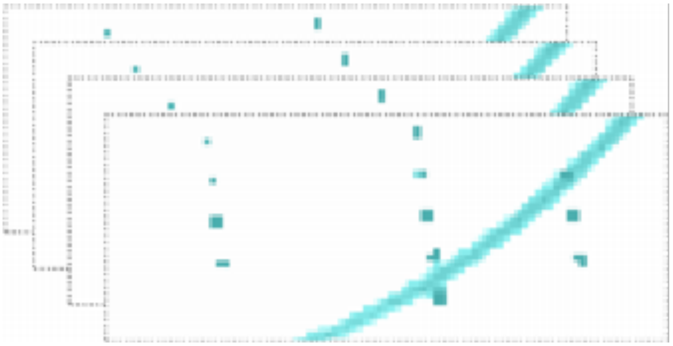
NSLS-II MX Data Management Plan



Herbert J. Bernstein, Dowling College

Talk for

“Partnering Data Collection and Reduction in the Beamline Environment”,
Harvard Medical School, 27 July 2012

Overview

- Data reduction and structure solution at beamlines is of increasing importance
- New sources and new detectors bring new challenges
- NSLS-II MX will be a BIGDATA problem and will need a well-designed data management plan both for beamline data process and subsequence processing.
- The issue of fraud in crystallography complicates the problem
- Some or all of raw data will need to be retained
- Handles, digital signatures and compression (lossless and lossy) will be needed. Data reduction to spots, structure factors or structures are examples of lossy compression.
- HDF5, NeXus, CBFlib and database access will need to be integrated

Beamline Data Reduction

- Beamline data reduction and structure solution “significantly accelerates the process of structure determination and on average minimizes the number of data sets and synchrotron time required for structure solution.” [W. Minor, M. Cymborowski, Z. Otwinowski and M. Chruszcz, “HKL-3000: the integration of data reduction and structure solution - from diffraction images to an initial model in minutes”, Acta Cryst. (2006). D62, 859-866]
- MX structure factors can be generated by an automaton (see, e.g. [G. Winter, “xia2: an expert system for macromolecular crystallography data reduction”, J. Appl. Cryst. (2010). 43, 186-190]
- Suggests that we are entering an era when the data delivered by a beamline may be background-removed spots, structure factors, or even solved structures.

Are raw images still needed?

NSLS-II MX

- NSLS-II: new state-of-the-art, medium-energy electron storage ring (3 billion electron-volts)
- designed to deliver world leading intensity and brightness
- will produce x-rays more than 10,000 times brighter than the current NSLS
- Advanced Beamlines for Biological Investigations with X-rays (ABBIX) are being built for NSLS-II, e.g.
  - Frontier Macromolecular Crystallography (FMX) beamline
  - Automated Macromolecular Crystallography (AMX) beamline
  - X-ray Scattering for Biology (LIX) beamline
- Some will have very high data fluxes and data rates, FMX 10^{13} photons/sec, AMX 2 x 10^{13} photons/sec
- Expect raw data rates of 9 gigabytes/sec = 72 gigabits/sec

A data management plan is needed

The Issue of Fraud in Crystallography

In late 2009, Dauter and Baker [Daut 10] expressed the shock of the Crystallographic community in discovering that a small, but significant, problem of fraud had developed in crystallographic structure determination and in an editorial in Acta Cryst. D wrote:

“The recent announcement made by the University of Alabama at Birmingham (USA) that a number of crystal structures produced in the laboratory of Dr H.M. Krishna Murthy will have to be removed from the Protein Data Bank and retracted from the literature, spread a shockwave among the macromolecular crystallography community. These structures were published over a period of seven years (1999 – 2006), and included such important proteins as dengue virus protease (1bep, 1df9, 2qid), complement component proteins (2hr0, 1g40, 1g44), vaccinia complement proteins (1rid, 1y8e), apolipoproteins (1i6l, 2ou1, 2a01), and Taq DNA polymerase (1cmw, 1bgx). In the past there have been cases in which structures determined by X-ray crystallography have had to be retracted because of errors in data interpretation or in the programs utilized in structure solution. These are understandable, as mistakes do occur. This time, however, it appears that the retracted structures were deliberately fabricated and there is no evidence that any experimental data were actually collected. ...”

Forensic Quality Data

Sound design of a data-management system for crystallographic data should allow for the possible future need to review carefully any given dataset of raw data to try to eliminate the possibility that it was fabricated. We refer to such raw data as "forensic quality" data.

The Murthy data made it clear that structure factors alone may not be sufficient for such forensic investigations.

The challenge for the data system is to provide for the necessary forensic audit trail with digital signatures where needed, without unnecessarily burdening the handling of those portions of the data stream for which such policies are not appropriate.

Why Preserve Data for Published Structures

"... Well.....Why publish data? Please let me offer some reasons:

- To enhance the reproducibility of a scientific experiment
- To verify or support the validity of deductions from an experiment
- To safeguard against error
- To allow other scholars to conduct further research based on experiments already conducted
- To allow reanalysis at a later date, especially to extract 'new' science as new techniques are developed
- To provide example materials for teaching and learning

Why Preserve Data for Published Structures (continued)

- To provide long-term preservation of experimental results and future access to them
- To permit systematic collection for comparative studies
- And, yes, [Better to] safeguard against fraud than is apparently the case at present

Also to (probably) comply with your funding agency's grant conditions:- Increasingly, funding agencies are requesting or requiring data management policies (including provision for retention and access) to be taken into account when awarding grants. ..."

[J. Helliwell. "Re: very informative - Trends in Data Fabrication". CCP4BB Archives, April 2012.]

Therefore …?

- Reducing data at the beamline is a good idea

- Reducing data at the beamline can be very helpful in determining what data to bring forward to publication

- But, more than the reduced data may be needed for structures that are brought forward to publication, and in some cases there may be scientific value or legal reasons to retain raw data even if it is not the basis for a publication

- A good data management plan will facilitate both reduction of the data and retention of raw data when appropriate

Data Management Plan Principles

- Know which data you need to keep and which data you don’t need to keep
  - Keep appropriate metadata on retention requirements
  - Hold dates, stakeholders
  - Keep the metadata with the data to avoid mistakes
- Know what data you have and where it is
  - Keep a database of the data you have and where it is
  - Also keep the retention metadata here
  - Keep digital signatures of all data
  - Use Handles (or DOIs) to identify data that may move
- Minimize the resources needed to keep the data you do need to keep.
  - Don’t keep data longer than you need to
  - Make the data you do keep as small as possible
- Use HDF5, NeXus, CBFlib and SQL

Minimize Resources

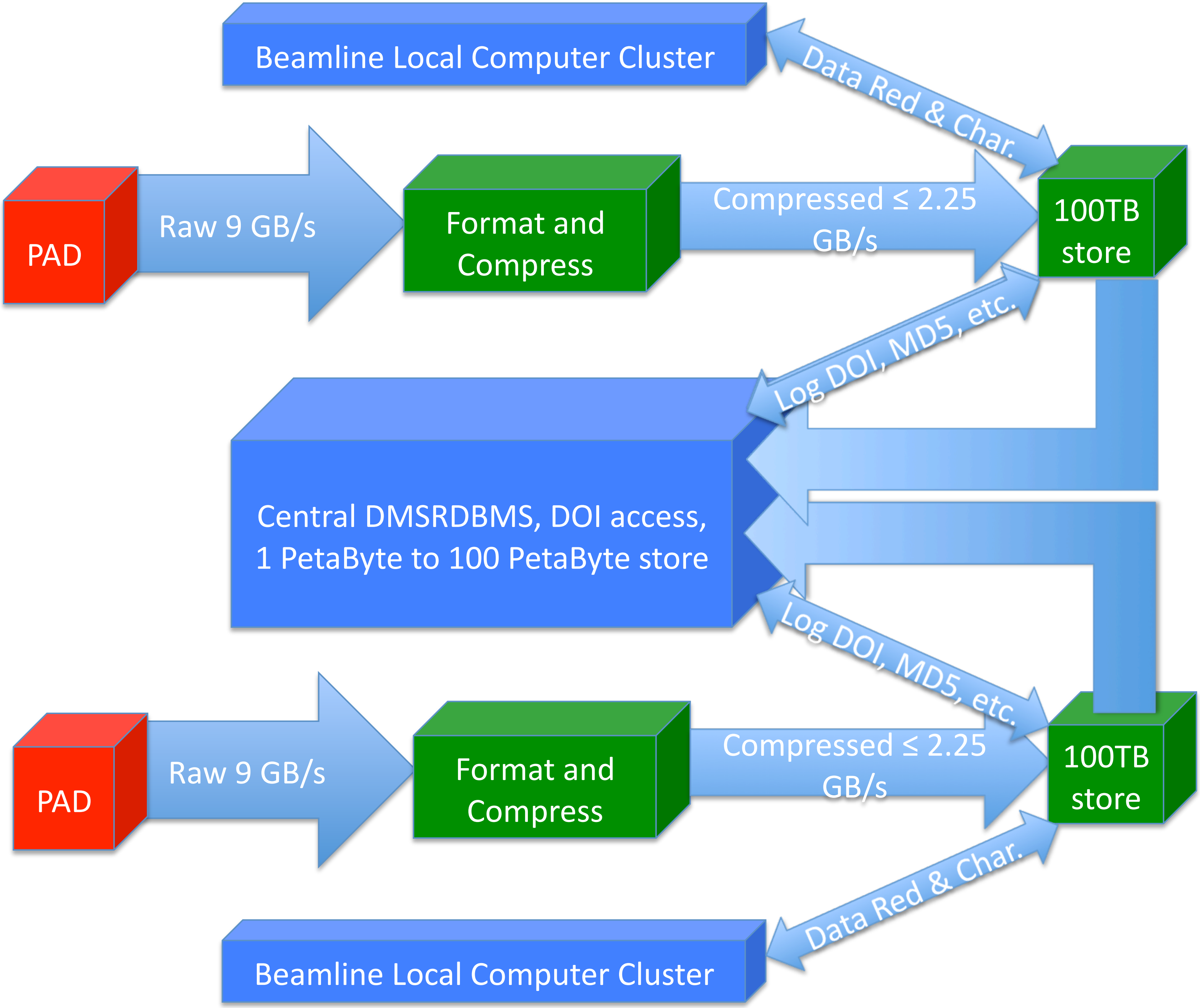
- Don’t keep data longer than you need to
  - Enforce the metadata-driven retention requirements
  - Do keep digital signatures of all collected data, even if the data itself has been handed off to the user and purged locally
- Make the data you do keep as small as possible
  - Use lossless compression (4:1 – 10:1) immediately
  - As per James Holton, “with 20:1 lossy compression of a crystallographic dataset “images will visually look pretty much like the originals, and generally give you very similar Rmerge, Rcryst, Rfree, I/sigma, anomalous differences, and all other statistics … Essentially, lossy compression is equivalent to adding noise to the images.” [J. Holton. “Image Compression”. CCP4BB Archives, November 2011]

Minimize Resources (continued)

- Wavelet compression can be lossless or lossy, essentially generalized spot-finding that retains some or all background. [J. Ferrer, M. Roth, and A. Antoniadis. “Data compression for diffraction patterns”. Acta Crystallographica Section D: Biological Crystallography, Vol. 54, No. 2, pp. 184–199,1998.
- Spot finding is a more drastic form of lossy compression (usually better than 1000:1) but may lose information on ice rings, split crystals, weak spots
- Structure factors, maps, etc are also lossy compressions

Plan Outline

- Software Elements
  - DMSRDBMS: data-management-system relational database with OSTI DOI interface
  - HDFCN: HDF5 data-management system integrated with CBFlib and NeXus interface
  - PADI: Interface software for pixel array detectors
- beamline: collect approximately 720 datasets per day using PADI and HDFCN, 47 terabytes uncompressed stored as approximately 12 terabytes compressed data with metadata and checksums
- from beamline to NSLS-II central repository via HDFCN, transfer recompressed data daily (2.4 – 7 TB/beamline) and update DOI location information
- central repository: retain data for appropriate retention period (minimum 1 week, typically 1– 6 months, in some cases multiple years)
- central repository: purge data on basis of retain/discard policies recorded in DMSRDBMS



Beamline Local Computer Cluster

Data Red & Char.

PAD

Raw 9 GB/s

Format and Compress

Compressed ≤ 2.25 GB/s

100TB store

Log DOI, MD5, etc.

Central DMSRDBMS, DOI access, 1 PetaByte to 100 PetaByte store

Log DOI, MD5, etc.

PAD

Raw 9 GB/s

Format and Compress

Compressed ≤ 2.25 GB/s

100TB store

Data Red & Char.

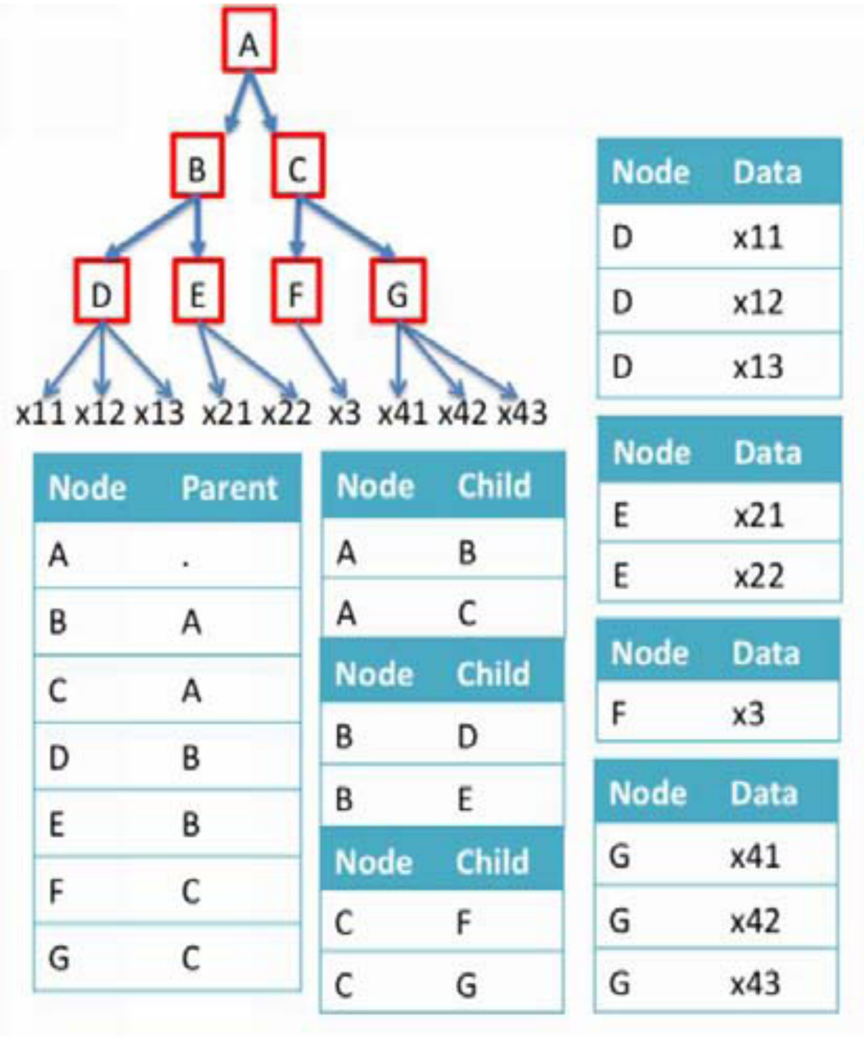
Beamline Local Computer Cluster

HDF5, NeXus, CBFlib and Database Access

- The Hierarchical Data Format Version 5 (HDF5) is a self-describing file format with a robust, well documented API capable of handling (and routinely used to handle) multi-gigabye files of data.
- NeXus is a tree-oriented view of HDF5 (and XML and HDF4) of importance in managing neutron and x-ray data. NeXus is a convenient thin layer over HDF5 that is widely accepted at many physics research centers, including at synchrotrons.
- The Crystallographic Binary File (CBF) format is a complementary format to the Crystallographic Information File (CIF), supporting efficient storage of large quantities of experimental data in a self-describing binary format
- HDF5 is tree-oriented. CBF is table-oriented, an essential approach for database access. We are combining CBFlib with an HDF5 backend to provide full database access to the metadata.

Converting to Tables

In order for information to be easily, reliably and efficiently searched, tables are more useful than trees, allowing the information to be loaded into a relational database management system [E. F. Codd. “A relational model of data for large shared data banks”. *Communications of the ACM, Vol. 13, No. 6, pp. 377 – 387, 1970.*].



Node	Parent
A	.
B	A
C	A
D	B
E	B
F	C
G	C

Node	Child
A	B
A	C

Node	Child
B	D
B	E

Node	Child
C	F
C	G

Node	Data
D	x11
D	x12
D	x13

Node	Data
E	x21
E	x22

Node	Data
F	x3

Node	Data
G	x41
G	x42
G	x43

Critical Issues

- Introduce new high performance compressions in HDF5 backend to CBFlib (nibble-offset compression for clean background PADs, lossless and lossy wavelet compression)
- Upgrade from MD5 to SHA2 and SHA3 digital signatures for images
- Map full NeXus tree to tables and add HDF5 support for relational table-based queries

Bibliography

[Bern 04] D. J. Bernstein. “Document IDs”. 2004. http://cr.yp.to/bib/documentid.html.

[Bern 05] H. J. Bernstein and A. P. Hammersley. “Specification of the Crystallographic Binary File

(CBF/imgCIF)”. In: S. R. Hall and B. McMahon, Eds., *International Tables For Crystallography,* Chap. 2.3, pp. 37 – 43, International Union of Crystallography, Springer, Dordrecht, NL, 2005.

[Bern 11a] D. Bernstein, N. Duif, T. Lange, P. Schwabe, and B. Y. Yang. “High-speed highsecurity signatures”. In: *Cryptographic Hardware and Embedded Systems– CHES 2011, pp. 124 – 142, Springer, 2011*

[Codd 70] E. F. Codd. “A relational model of data for large shared data banks”. *Communications of the ACM, Vol. 13, No. 6, pp. 377 – 387, 1970.*

[Daut 10] Z. Dauter and E. N. Baker. “Editorial: Black sheep among the flock of protein structures”. *Acta Cryst., Vol. D66, No. 1, p. 1, 2010.*

[Elli 05] P. J. Ellis and H. J. Bernstein. *Definition and Exchange of Crystallographic Data, International Tables For Crystallography, Chap. CBFlib: an ANSI C library for manipulating* image data, pp. 544 – 556. International Union of Crystallography, Springer, Dordrecht, NL, 2005.

Bibliography (continued)

[Ferr 98] J. Ferrer, M. Roth, and A. Antoniadis. “Data compression for diffraction patterns”. *Acta Crystallographica Section D: Biological Crystallography, Vol. 54, No. 2, pp. 184–199,* 1998.

[Filg 01] U. Filges. “The new NeXus API based on HDF5”. In: *VITESS Workshop Berlin, 25 – 27 June 2001, 2001.*

[Gotz 10] A. G¨otz, V. Sol´e, C. Madonna, and A. F. Maydew. “ELISA VEDAC Workshop Report, Workshop Title: HDF5 as hyperspectral data exchange and analysis format, Grenoble, January 11th to January 13th, 2010.”. January 2010. http://vedac.esrf.eu/public-discussions/hdf5-workshop/workshop-report.

[Hall 05] S. R. Hall and B. McMahon. *Definition and Exchange of Crystallographic Data, International Tables For Crystallography, Chap.Genesis of the Crystallographic Information* File. Vol. G, Springer, Dordrecht, 2005.

[Hall 91] S. R. Hall, F. H. Allen, and I. D. Brown. “The Crystallographic Information File (CIF): a new standard archive file for crystallography”. *Acta Crystallographica Section A: Foundations of Crystallography, Vol. 47, No. 6, pp. 655 – 685, 1991.*

Bibliography (continued)

[Haml 12] R. Hamlin, T. Hontz, and C. Nielsen. “The New Dual Mode Pixel Array Detector”. In: *Meeting of the American Crystallographic Association, Boston, MA, 28 July – 1 August2012, American Crytallographic Association, 2012. Abstract 11.01.1151, to appear.*

[Hell 12] J. Helliwell. “Re: very informative - Trends in Data Fabrication”. CCP4BB Archives, April 2012. https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1204&L= ccp4bb&F=&S=&P=122036.

[Hend 12] W. A. Hendrickson. “NSLS-II - Status of the Life Sciences Program”. Tech. Rep., Brookhaven National Laboratory, X6A Science Advisory Committee, February 2012. protein.nsls.bnl.gov/mediawiki/images/e/e3/Hendrickson_2012.pdf.

[Holt 11] J. Holton. “Image Compression”. CCP4BB Archives, November 2011. https://www. jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1111&L=ccp4bb&F=&S=&P=71742

[Lang 04] M. Langston and J. Tyler. “Linking to journal articles in an online teaching environment: The persistent link, DOI, and OpenURL”. *The Internet and higher education, Vol. 7,* No. 1, pp. 51–58, 2004.

Bibliography (continued)

[Ozak 07] S. Ozaki, J. Bengtsson, S. Kramer, S. Krinsky, and V. Litvinenko. "Philosophy for NSLS-II design with sub-nanometer horizontal emittance". In: *Particle Accelerator Conference, 2007. PAC. IEEE, pp. 77–79, IEEE, 2007.*

[Pren 10] B. Preneel. "The first 30 years of cryptographic hash functions and the NIST SHA-3 competition". *Topics in Cryptology-CT-RSA 2010, pp. 1 – 14, 2010.*

[Rew 04] R. Rew, B. Ucar, and E. Hartnett. "Merging netCDF and HDF5". In: *20th Int. Conf. on Interactive Information and Processing Systems, 2004.*

[Rive 78] R. Rivest, A. Shamir, and L. Adleman. "A method for obtaining digital signatures and public-key cryptosystems". *Communications of the ACM, Vol. 21, No. 2, pp. 120–126,* 1978.

Acknowledgements

BNL PXRR Group

- Robert M. Sweet
- Dieter Schneider
- Howard Robinson
- John Skinner
- Matt Cowan
- Leonid Flaks

Dowling College ARCiB Lab Group

- Herbert J. Bernstein
- Ming Li
- Mojgan Asadi
- Keti Bardhi
- Kostandina Bardhi
- Limone Rosa

Frances C. Bernstein

Work funded in part by NIH, DOE, NSF