Computing Issues and SLD Data Archive

SLD Collaboration Meeting

Kirkwood, Summer Solstice, 2000

Tony Johnson (tonyj@slac.stanford.edu)

Computing Status

- VMS Cluster is very stable
  - But getting old
  - SLACAX remains on 24x7 support
- Currently expected to be closed down:
  - When "*SLD finishes its off-line data analysis*"
  - Currently scheduled for end of 2001
- Why shut it down at all?
  - SCS Support Manpower (25% FTE)
  - Maintenance costs (currently $1000/month)
  - Space in computer room


Kingsburg
20


Sun

Tape Access

- Currently have 10,000 slots in silo
  - We will have to half this by the end of year
    - Should be no problem
- SLD currently uses 1GB tapes
    - By end of 2001 SCS wants to get rid of all old drives and move all old tapes out of silo
    - Switch to new 60GB tapes
- When VMS shuts down all access to SLD style tapes will be terminated
  - VMS cannot read new tapes

Consequences

- When VMS shuts down
  - Recon and MC will no longer be usable
  - IDA, analysis utilities will no longer be usable
  - Existing tapes will not be readable
- If we want to maintain access to any of our data we need to act before VMS shuts down.

Archiving SLD Data

- Software issues
- Hardware issues
- Sociological/Philosophical issues

What is LEP's plan?

- LEPC developed plan for archiving LEP data
  - Make data accessible for 30+ years
  - Archive 4 experiments in common format
    - 4-vectors
    - Archive using "LHC technology" (Objectivity?)
    - Allow public access to data
    - http://lhcb.cern.ch/~cattanem/Aleph/lep-archive980423.ppt
- Appears to have been rejected by experiments
  - Did not think common format was practical
    - 4 vectors now useful
  - Could not agree on usefulness/desirability of making data publicly available

Current LEP effort

- Only ongoing effort I found was:
  - Aleph – copying data to objectivity
    - Seems to be aimed more at learning about/experimenting with Objectivity than long term archive
    - `http://alephwww.cern.ch/~disserto/objty`

Archiving Mini DST

- What format to store data in
  - Jazelle, Paw, Root, Objectivity
    - All formats undocumented
      - Can only be read by the program that wrote them
      - Can become obsolete
        - If program becomes obsolete
        - If new version will not read old data
- Alternatives?

What do we want to save?

- MiniDST
  - R17 data + MC
- What else needed for important analyses?
  - 120Hz data needed for A_{lr}
  - Other?
    - Private N-tuples
- Raw Data
  - Is there much point if we can’t use them?

XML

- XML is generalized markup language
- “Hot format” for data storage/transfer
- ASCII format
  - Anyone can read it (even Fortran program)
- Very large (can use compression)
- Can loose precision
- Slow
  - If no one ever uses data it is unlikely to be correct.

SIO

- Very simple format developed by Tony Waite for Linear Collider Studies
  - Based on XDR
    - Industry standard
    - Very simple format, well documented
  - SIO Adds support for
    - Event records
    - Pointers
  - SIO also very simple and documented
  - Much faster to read than XML

MP3

- Advantages
  - Collect large royalty checks from new age music fans
    - "Song of the Subatomic Particles"
  - Data will be automatically distributed and archived for us by college students using Napster
- MP3 players have limited analysis tools

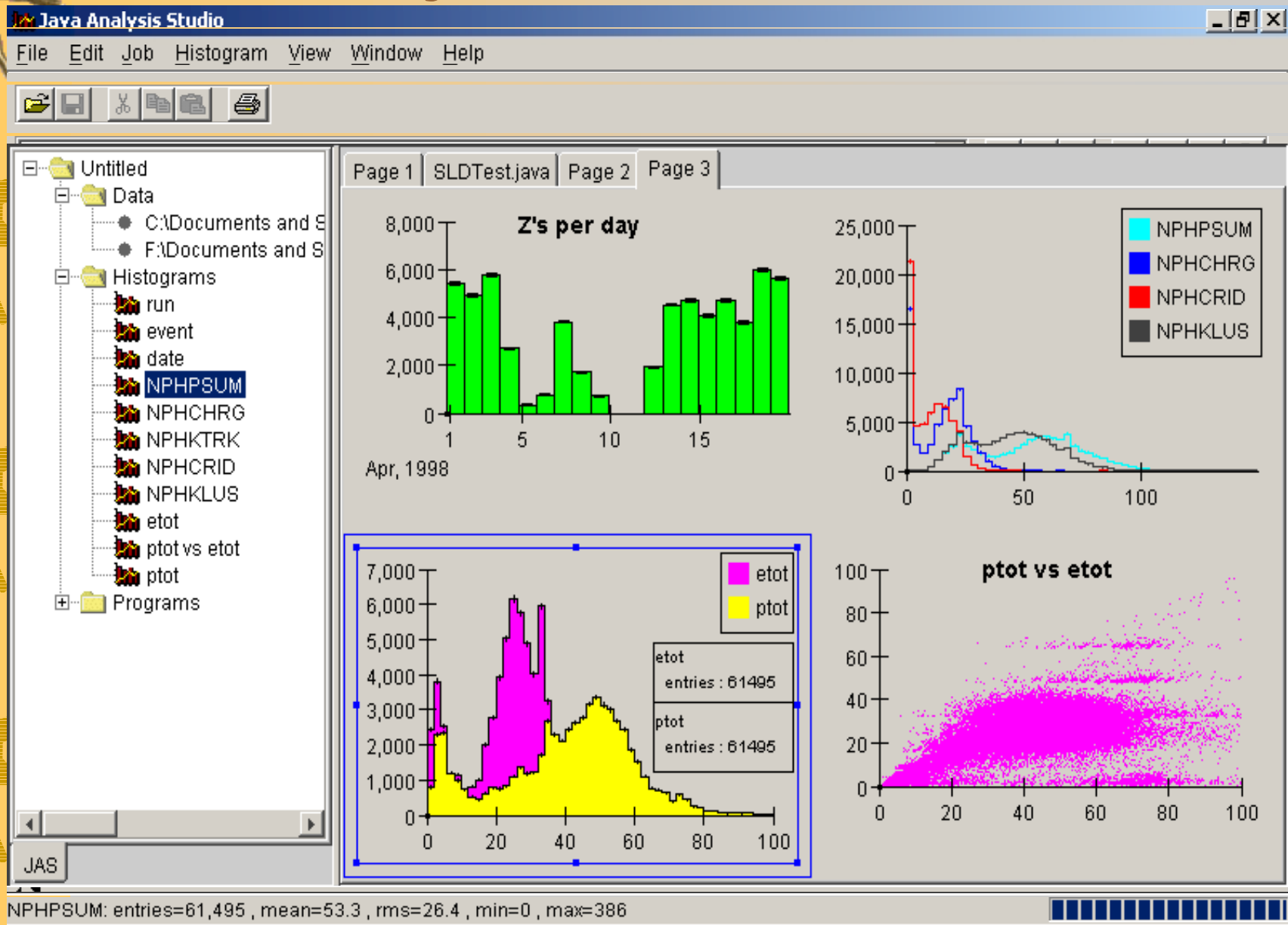
Winner is SIO?

- Started to write program to convert Jazelle Mini DST files to SIO
  - How to write it?
    - Fortran? – too boring
    - C++? -- too hard
    - Java? – sounds fun
      - Use Jazelle to read data and write out using Java?
        - Mixed Java + Fortran? -- Too Messy
      - Use Java to read Jazelle files
        - Turns out to be not too hard (read only, minidst only)
        - Takes < 500 lines of code

Winner is Jazelle?

- No conversion/validation of data necessary
  - We already know Jazelle data is as good as we can get it.
  - Even if Java becomes obsolete, or you just don't want to use Java, the (well documented) Java program shows how to read the data in any language you like, or you can use Java program to convert data to any other format you like.
- Can use JAS to analyze SLD data
- Could feed SLD data in Linear Collider analysis
  - Useful?

SLD analysis with JAS



Physical Access to Data

- Existing tapes become unreadable after VMS shutdown
  - Our existing tapes hold 1GB
- New tapes SCS is purchasing for BaBar hold 60GB
- SCS prepared to offer some assistance in copying data from old tapes to new
  - Want us to use HPSM mechanism
  - 60GB tapes will themselves be obsolete before long
    - Disk cheaper than tape in 5 years?

How much data do we have?

- Sample Data Set Sizes

Data Type	Data	MC
MiniDST	15 tapes	211 tapes (includes some junk)
S120	153 tapes	
RAW (ACQ)	8000 tapes	

- Also full recon data, Raw Data strips

What to archive?

- MiniDST can be stored on unix disk
  - Data + MC
  - Could also store on DVDs
    - One per institute??
- S120, HAD, Bhabha raw data can fairly easily be copied to 60GB tapes?
- Is there any point copying the ACQ tapes?

What else should we keep

- Code (all of DUCS?)
  - Can't run it but we can at least read it.
- Web Pages
  - Active web pages won't work anymore
  - Documentation pages could be moved to Unix or NT
- Distilled knowledge?
- What else?

What do we want to do with the data?

- Keep it “just in case”
  - Analyses are not finished before VMS shutdown
  - Some new interest in the future
- Make it publicly available?
  - Is there any point, could anyone use it?
- Educational Web Site?
  - Target to undergrads?
    - Needs to be interesting
    - Needs to be challenging
    - Doesn’t need to be a real SLD analysis

Conclusions

- VMS system will run to at least end of 2001
- Access to Mini DST (but not IDA, analysis tools) will still be possible after that
  - I recommend keeping the data in Jazelle format
  - Maybe you want to try the alternative access now?
- Need to carefully think about what we want to keep long term, and what we want to do with it?
  - This meeting would be a good time to think about it