Fermi National Accelerator Laboratory

FERMILAB-Conf-98/335-E
D0

Assurance of Data Integrity in Petabyte Data Samples

Heidi Schellman
For the D0 Collaboration Data Access Group

Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510

December 1998

Published Proceedings of *Computing in High Energy Physics (CHEP98)*,
Chicago, Illinois, August 31 - September 4, 1998

Operated by Universities Research Association Inc. under Contract No. DE-AC02-76CH03000 with the United States Department of Energy

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Distribution

Approved for public release; further dissemination unlimited.

Copyright Notification

Assurance of Data Integrity in Petabyte Data Samples

Heidi Schellman for the D0 Collaboration Data Access Group

Northwestern University, Evanston, IL 60208

Abstract. We present a method for clustering data from high energy physics experiments on physical media. Data clustering based upon physics information can lead to large gains in access speed. However such clustering also increases vulnerability to data loss as any loss of a physical data store removes a data sample which has special physics characteristics. Measurements made with samples biased by losses of clustered data must be corrected for the effects of the loss. We discuss several methods for performing such corrections.

INTRODUCTION

In the year 2000, the D0 collider experiment at Fermilab will begin data taking. This experiment is expected to write data at a rate of approximately 50 Hz with data rates of 10-20 MBytes/second. The data sample at the end of three years is expected to approach 1 petabyte in size with over 1 billion records of individual high energy interactions. These data must be stored on tape, processed through reconstruction algorithms and summarized for further study by the collaboration. [1]

D0 is a multi-purpose detector and the resulting data sets will be quite diverse. It is expected that several hundred physicists will access subsamples of the data, with individual searches using between 1-30% of the total. As any individual use of the data only accesses a small subsample, the collaboration is considering data clustering, in which data most likely to be accessed together are clustered on tape. Such clustering can improve access times by large factors but can also introduce biases into query results in the not unlikely event of hardware or software failures. In unclustered data, any loss is distributed randomly across all data types and has negligible effect on statistical results derived from those data. Clustered data can also provide unbiased results even in the presence of significant data losses if sufficient error recovery and monitoring are present.

I DATA CLUSTERING

Data Clustering can best be explained by an example – imagine an experiment which has 4 trigger types:

- E - a high p_T electron
- MU - a high p_T muon
- MET - large missing E_T
- SVTX - a separated vertex

Events from such an experiment can have $2^N - 1$ possible combinations of triggers where N is the number of trigger types. In this example there are 15 possible trigger combinations.

A Data storage possibilities

Consider four possible methods of storing these data:

1. No streaming - Write all triggers out to one stream and read the whole stream for each analysis. This makes access times for any individual sample very large.

2. Inclusive streaming - Write all triggers for a given analysis out to a special stream, any events which would be used in two or more analyses would be written out twice. This minimizes access time but raises the media cost.

3. Physical clustering - Write each of the 15 possible trigger combinations to its own physical stream. Read several physical streams to make a single logical stream. This minimizes both media cost and access time but introduces complexity and vulnerability to data loss.

4. Database - Write each event once and then use a database to optimize access. This method allows full optimization of both access time and media cost. The complexity is handled by commercial software but vulnerability to data loss is still a problem.

For the large petabyte raw and reconstructed data samples, the D0 collaboration has chosen option 3, the simpler data clustering method, instead of a full database implementation, mainly because commercial solutions capable of handling data samples of this size will not be available at reasonable cost on the D0 time scale.

B Data analysis using clustered data

Data analysis in the clustered scheme is bases on logical streams, which can be defined dynamically based on the triggers needed for a given data analysis. In this particular data clustering example, a $W \to e\nu$ analysis would use the four trigger combinations with both E and MET while a B$\to$ $\mu + X$ analysis would use all four combinations with both MU and SVTX and a top analysis might use only the E·MU·SVTX·MET combination.

We have done studies [2] of clustering using trigger samples from the previous collider run. These studies indicate that traditional inclusive streaming would create an overlap overhead of 30-50%. On the other hand, if data clustering is used, there is no overlap overhead and it was possible to access all W(E+MET) triggers by reading less than 3000 of the 100,000 sample events. In a real data analysis this would correspond to 150 instead of 5000 tape mounts. Similar results hold for other trial data analyses.

For a real implementation, our studies indicate that a simple choice of $2^N - 1$ streams for N triggers is too naive. In practice we will divide the 128 possible triggers into $N \approx 7$ trigger groups, find the $2^N - 1$ possible combinations and merge small combinations which are similar until we have of order 10-15 physical streams. This algorithm is easy to automate. [4]

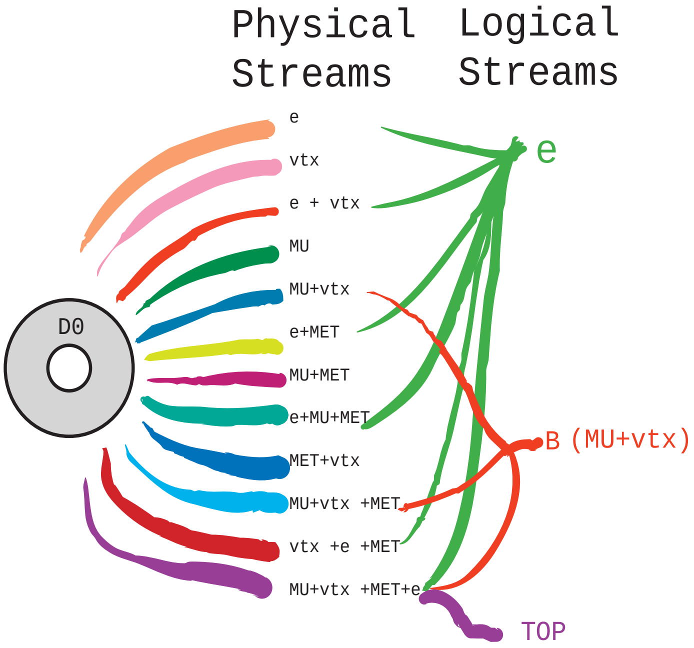

FIGURE 1. Illustration of data clustering. The data are written out to many exclusive physical streams which are then combined into logical streams (equivalent to inclusive streams).

II DATA INTEGRITY

Clustering data according to trigger type or access pattern can provide significant time and space savings. However, using physics criteria to cluster data on physical media has its perils. Any loss of data will be introduce a physics bias because the data lost are more similar to each other than to other data. In the simple B physics example given in section 1, loss of one file (perhaps a file which has (MU·SVTX·$\overline{\mathrm{E}}$·MET) triggers in it) might introduce a bias into a measurement of B fragmentation by selectively removing events with large higher p_T for the recoil system. If the effect of this loss is large and cannot be simulated accurately, the most conservative solution is to eliminate all live-time corresponding to that file from the data sample. In the example given the logical MU·STVX stream is made of the 4 physical streams which have both MU and SVTX triggers. Loss of one file in one stream has resulted in similar losses in all 4 streams.

This result applies to any system, including databases, where data are clustered based on physics criteria; permanent loss of data leads to biases which can be avoided by removing yet more data from the sample – the more efficient the clustering, the larger the potential loss.

In fact, if data are written out to files of constant size, the multiplication factor is approximately the number of physical streams used in the analysis. Streams with large amounts of data will lose a smaller fraction of livetime from one tape loss but

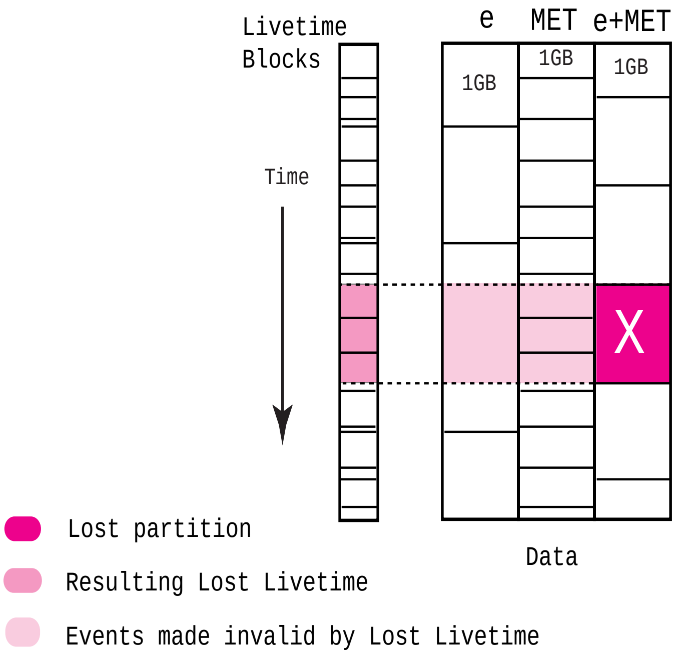

FIGURE 2. Data vs Time for several physical streams. Loss of one file results in losses of several other files.

have more tapes to lose while streams with small amounts of data have fewer tapes to lose but each loss removes more livetime.

A livetime removal algorithm is reasonably easy to automate, a valid livetime list is maintained for each analysis and updated when files are lost or found.

In some cases such draconian removal of all data during suspect livetime is not necessary. It is possible to estimate the effects of a lost file using the remaining data or simulation and make a correction. This method is less easy to automate.

III CONCLUSIONS

We have presented a method for speeding access to large data sample in which the data are clustered based on anticipated access patterns. Preliminary studies indicate that this physical clustering method can produce large gains in access speeds but that such optimization increases the vulnerability of the data sample to data loss. Both livetime removal and data-based correction methods can compensate for these biases.

REFERENCES

1. D0 Note Number: 003465
   Title: Requirements for the Sequential Access Model Data Access System
   Author(s): Jon Bakken, Mike Diesburg, Dorota Genser, Lee Lueking, Frank Nagy, Don Petravick, Ruth Pordes, Heidi Schellman, Marilyn Schweitzer, Igor Terekhov, Matt Vranicar, Vicky White
   Date: 6/15/98
2. D0 Note Number: 003326
   Title: Study of Possible Run II Streaming Scenarios Using a Run 1 Data Sample
   Author(s): Schellman H., Bertram I.
   Date: 10/21/97
3. D0 Note Number: 003313
   Title: New Phenomena Comments on the D0 Run II data Access
   Author(s): Bantly J., Boehnlein A., Hobbs J.
   Date: 9/30/97
4. D0 Note Number: 003523
   Title: Summary of the Luminosity Workshop - September 17th-18, 1998
   Author(s): H. Schellman, R. Partridge - convenors C.C. Miao, Ph. Laurens, D. Edmunds, G. Brooijman, L. Paterno, F. Bartlett, S. Fuess, L. Lueking, V. White, H. Melanson, J. Yu, I. Bertram, J. Krane etc. - participants
   Date: 10/5/98