Proposal for Thumbnail Contents

Ulrich Heintz

for the Data Tier Committee[i]

Version 3

December 6, 2000

In Run I, D0 used a microDST format The following three principles minimized its size:

- Only selected words were stored (not whole ZEBRA banks)
- The ZEBRA overhead was minimized (the microDST was stored as a single ZEBRA bank and converted into the standard DST format at analysis time.
- Storage of redundant information was avoided where possible. For example, instead of pT, px, py, pz, E, phi, eta, theta, only pT, phi, and eta were stored, the others were recalculated at analysis time.

The expansion of the microDST into the standard DST format was performed by a software package, entirely transparent to the user.

For Run II, the Data Tier Committee proposed a data format called thumbnail. Data in this format should always be resident on disk. It serves two purposes:

- Creation of event lists based on a set of analysis cuts that can be used to access DSTs for the selected events.
- Some basic analysis capability.

The goal is to keep the size of the thumbnail as small as possible, with a tentative target of 5kB/event and a hard upper limit of 10 kB/event.

The following is a proposal for the contents of this thumbnail. This proposal summarizes a series of presentations of the object ID groups to the Data Tier Committee. For each type of object a list of quantities is given. For each quantity, the number of values saved and the estimated number of bits required for each value are given. Some quantities are stored once per event, some once per algorithm, and some once per object, as indicated in the last three columns. At this time we foresee the following object types:

- Electromagnetic Clusters
    - All electrons (soft and standard)
    - All isolated photons
- Muons
- Taus
- Jets
- Missing pT
- Tracks

    - All tracks found by the reconstruction
    - More detailed information saved for isolated tracks
- Vertices
- Preshower clusters
    - All preshower clusters associated with electrons, taus, and isolated tracks
- Trigger information

For each object type, we have to keep a list of chunk id's and rcp id's. The first list tells us what input data was used to make the chunk and the second list tells what algorithms were used.

The information stored on the thumbnail allows rerunning the following reconstruction tasks. Vertex finding can be redone, so that algorithms to find displaced vertices can be optimized using the thumbnail. Jet energy corrections can be applied to the jet objects on the thumbnail. Missing pT can be determined to good approximation with respect to any arbitrary z-vertex position.

It has been requested that the Monte Carlo information should be compressed for the thumbnail as well. We would like to defer the development of a specific proposal for this. As a default solution the MC chunk could be stored on the thumbnail as is.

Open Questions

The format to store the thumbnail in and how much overhead will be associated with this format are not decided. We propose to use D0OM to save the thumbnail. This implies that we expect the TMB to be a single chunk. We will use the event data model and the I/O can be whatever interface exists to D0OM. The reason for this choice is twofold: (1) we can implement the packing of words into as few (8, 16, 32) bits as needed and (2) we can make the analysis code independent of the file format (DST or thumbnail) by providing a transparent unpacking routine, much like it was done in Run I.

Whatever format is ultimately adopted for the thumbnail, we propose to create a TMB chunk that contains the data of the entire thumbnail. This chunk will be kept as part of the DST data format.

In many cases, pointers from an object reference other objects. It has not been decided how to implement such pointers. These could be implemented with 8-bit words (as currently specified in most objects) or as bit masks with as many bits as necessary for the event. We could also use both approaches (whichever uses less bits in any given case), with a control bit that indicates which was used.

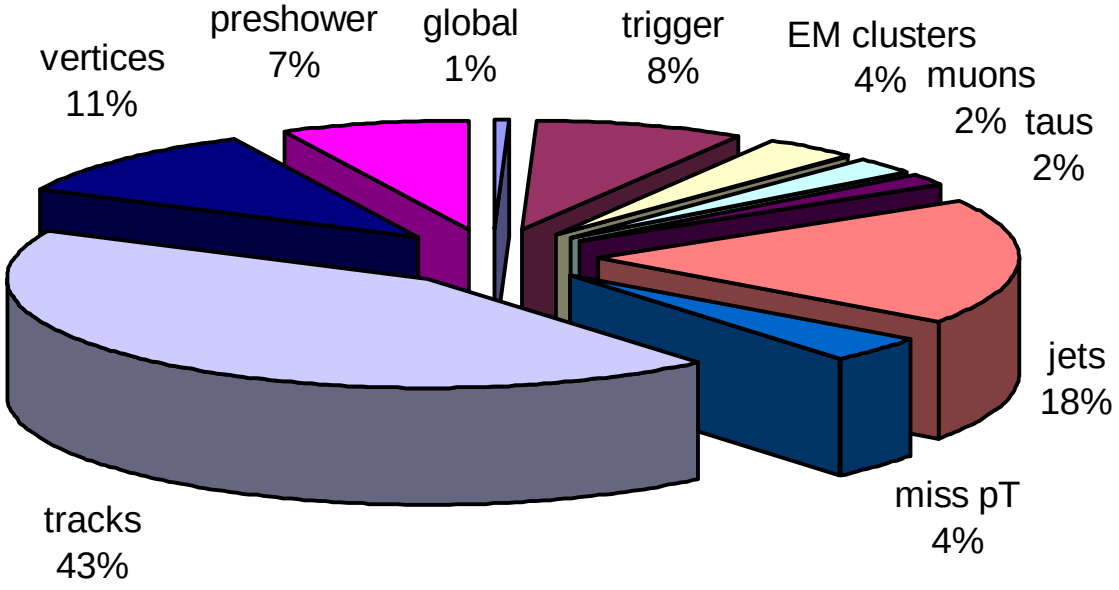
We do not have any quantitative studies that indicate that all error matrices can be saved in 16-bit precision as currently assumed. We may have to use 32 bits to maintain the precision necessary for refitting vertices or performing kinematic fits. This is an important question that merits careful study. If all error matrix elements were saved in 32-bit precision, the size of the thumbnail would increase by over 2 kB/event.

Thumbnail Size

This table summarizes the size of the data contained in the proposed thumbnail. The following tables list the detailed contents for each object type.

object type	bytes/event	bytes/algo	algorithm	bytes/object	objects	total bytes
global	48	0	1	0	1	48
trigger	569	0	1	0	0	569
EM clusters	0	0	1	128	2	256
muons	0	0	1	90	2	180
taus	0	0	1	69	2	139
jets	0	42	4	73	4	1334
miss pT	6	0	1	280	1	286
tracks	0	0	1	39	80	3128
vertices	0	0	1	38	20	760
preshower	0	0	1	37	14	520
total						7219

The numbers of em clusters is based on Run I all stream micro DST. The number of muons was quoted by the muon ID group. The number of taus is a guess. The number of jets is based on the Run I all stream microDST. The jets group guessed 5. The number of tracks are based on ttbar events. The number of vertices are also based on ttbar events. The number of preshower clusters assume 2/electron, 4/tau, and 1/isolated track with 2 isolated tracks/event.



Global

quantity	values	bits	bits/event
run number	1	32	32
event number	1	32	32
turn number	1	32	32
bunch number	1	8	8
time	1	32	32
luminosity block number	1	32	32
luminosity block number status	1	4	4
instantaneous luminosity	1	16	16
reco version	1	16	16
number of objects	8	8	64
pointer to primary vertex	1	8	8
L0 information	1	32	32
detector condition flags	32	1	32
accelerator condition flags	32	1	32
magnet on/off	4	1	4
magnet polarities	4	1	4
total(bits)			380
total(bytes)			48

The luminosity block number and status are needed to calculate integrated luminosity.

Number of objects stores how many objects of each type this event contains.

The primary vertex is identified using only the tracks. No other object information enters into the decision. The vertex with at least 3 associated tracks that contains the track with the largest p_T is selected as the primary vertex. It is used to derive the momentum vectors of electromagnetic showers and jets.[ii]

The detector and accelerator condition flags are placeholders for the quantities that are saved in the event header.

Trigger[iii]

quantity		values	bits	bits/event
trigger name		8	128	1024
L2 trigger bit		8	1	8
L3 trigger bit		8	1	8
"and/or" trigger mask		1	256	256
L1 objects	muons	1	32	32
	EM reference sets	1	384	384
	jet reference sets	1	384	384
	large tile reference sets	1	96	96
	missing pT	1	8	8
L2 objects	electrons/photons	2	80	160
	muons	2	80	160
	taus	2	80	160
	jets	4	64	256

	missing pT	2	24	48
L3 objects	electrons/photons	2	80	160
	muons	2	80	160
	taus	2	80	160
	jets	4	256	1024
	missing pT	2	32	64
total(bits)				4552
total(bytes)				569

A trigger is tagged by a name (not a bit). This name is unique for a L1-L2-L3 combination and can be found in the database. The name is limited to 16 characters. The trigger names identify all L1 triggers that fired for this event. For each of these triggers, the status of L2 and L3 is also retained.

For each Level 1 reference set (4 em, 4 jet, 2 large tile sets) η,φ,pT for up to 4 objects are saved.

Track and preshower objects are too numerous and will not be stored.

Electromagnetic Clusters[iv]

quantity	values	bits	bits/object
charge*pT	1	16	16
phi	1	16	16
eta	1	16	16
error matrix	6	16	96
fit chi squared	1	16	16
number of degrees of freedom	1	8	8
pointers to tracks	2	8	16
pointers to preshower	2	8	16
isolation fraction	1	8	8
core energy	1	16	16
likelihood	1	16	16
H-matrix chi squared	1	16	16
phi of cluster	1	16	16
r of cluster	1	16	16
z of cluster	1	16	16
energy in FH2,3,4,CH	1	16	16
transverse shape	5	16	80
layer energy	6	16	96
<phi>,<z> weighted center of gravity in each layer	12	16	192
<phi^2>,<z^2> in each layer	12	16	192
sum of weights in each layer	6	16	96
number of cells in each layer	6	8	48
algorithm	1	8	8
total(bits)			1024
total(bytes)			128

There are two distinct algorithms for finding electron candidates and one for photon candidates. By definition photon candidates are isolated. All the em clusters found will be

treated on the same footing with an internal label (RCP id) identifying the algorithm used. The first electron algorithm starts from tracks, which are then associated with clusters in calorimeter and preshowers. The second starts from clusters and to avoid duplication only candidates not found with the first method are kept. Only one clustering algorithm will be used (whichever turns out to be most efficient). The construction of an em cluster is kept as a separate step from the algorithm used to find em clusters, so efficiency and resolution are decoupled.

The momentum of photons is calculated using the primary vertex (see global object) and the em cluster energy. The direction of an electron is determined by the track and the energy by the em cluster for those in which a cluster found beforehand was matched to a track, but by the track if the cluster had to be constructed using track information.

Isolation fraction is defined as $(E_{tot}(0.4)-E_{em}(0.2))/E_{em}(0.2)$ as in Run 1. Core energy is $E_{em}(0.2)$. The energy in the isolation cone with R=0.4 can be found using these two quantities.

The 5 transverse shape quantities are at present the energy of the 5 highest energy cells in EM3. This was found to be very effective for separating photons from jets.

Track match quality can be calculated from the information in the thumbnail.

The inputs for the likelihood calculation are Hmatrix, E/p, em fraction, and track match quality. These are available on the thumbnail. The Hmatrix chi squared may be based on the 41x41 matrix from Run 1 or on a new 10x10 matrix. In either case the matrix will not be stored on the thumbnail.

Muons[v]

quantity	values	bits	bits/object
global track parameters	5	16	80
error matrix	15	16	240
charge*pT from muon system	1	16	16
BC bend coordinate at toroid midplane	1	16	16
BC slope in bend view	1	16	16
BC nonbend coordinate at toroid midplane	1	8	8
BC slope in nonbend view	1	8	8
A bend coordinate at toroid midplane	1	16	16
A slope in bend view	1	16	16
A nonbend coordinate at toroid midplane	1	8	8
A slope in nonbend view	1	8	8
pointers to tracks	2	8	16
pointers to best vertices for tracks	2	8	16
quality	1	32	32
hit layers in fit	1	40	40
expected hit layers	1	40	40
A layer hits within 3 sigma	1	6	6
B layer hits within 3 sigma	1	5	5
C layer hits within 3 sigma	1	5	5

chi squared from global fit	1	8	8
chi squared for muon system	1	8	8
3 scintillator times in 1 ns bins	1	32	32
velocity flag	1	8	8
tracks with dR<0.5	1	8	8
momentum of tracks with dR<0.7	1	8	8
cal energy along nominal track	1	8	8
cal energy along best track	1	8	8
energy in last layer of calorimeter	1	8	8
energy in 3x3 window	1	8	8
energy in 5x5 window	1	8	8
energy in preshower	1	8	8
zvertex from MTC/error	1	8	8
total(bits)			720
total(bytes)			90

Muons require at least:

1. Gtrack with pT>1.5 GeV) and one of (Cal, A, BC), where A and BC denote at least 1 hit (wire or scintillator) inside and outside toroid, respectively.
2. BC segment + vertex + Cal
3. A segment + vertex + Cal + gap in central tracking

The quality word includes: Track quality, low ∫Bdl flag, points to odd vertex, thin toroid, part of J/psi or Z, hits on opposite side...(to be defined).

The hit layer bit mask is defined as follows:

SMT/CFT	FH,CH	Muon A	Muon BC (wire & scintillator)
22	5	5	8

Taus[vi]

quantity	values	bits	bits/object
charge (-1,0,+1)	1	2	2
pT	1	16	16
phi	1	16	16
eta	1	16	16
width	1	16	16
profile	1	16	16
fraction of energy in EM1	1	16	16
fraction of energy in EM2	1	16	16
fraction of energy in EM3	1	16	16
fraction of energy in EM4	1	16	16
fraction of energy in FH	1	16	16
fraction of energy in ICD/MG	1	16	16
fraction of energy in CH	1	16	16
hot cell ratio	1	16	16
fraction of energy in 3x3 towers	1	16	16
fraction of energy in 5x5 towers	1	16	16

energy of calorimeter cluster	1	16	16
center of gravity of calorimeter cluster (r,phi,z)	3	16	48
tracks in 10 degree cone	1	8	8
tracks in 20 degree cone	1	8	8
tracks in 30 degree cone	1	8	8
number of tracks	1	8	8
number of tracks in R<0.4 cone	1	8	8
track isolation	1	16	16
track rms	1	16	16
pointer to track 1	1	8	8
pointer to track 2	1	8	8
pointer to track 3	1	8	8
pointer to track 4	1	8	8
pointer to track 5	1	8	8
pointer to best vertex	1	8	8
pointer to preshower cluster	4	8	32
pointers to jets	4	8	32
PDE variable	1	8	8
likelihood	1	8	8
quality	1	32	32
Fisher variable	1	8	8
neural network	1	8	8
total(bits)			554
total(bytes)			69

Jets[vii] (including b/c tagging information[viii])

quantity	values	bits	bits/object	bits/algorithm
momentum	4	16	64	0
scalar ET	1	16	16	0
(EM/ICR/CH) fraction	3	8	24	0
FH 1,2,3 fraction	1	8	8	0
number of tracks	1	5	5	0
pointer to tracks	10	8	80	0
hot fraction	1	8	8	0
maximum cell ET	1	16	16	0
quality	1	32	32	0
eta/phi width	1	16	16	0
split/merge flag	1	10	10	0
energy before split/merge	1	16	16	0
charge	1	8	8	0
center of gravity of calorimeter cluster (r,phi,z)	3	16	48	0
z vertex projection	1	16	16	0
error on z vertex projection	1	16	16	0
preshower energy	1	16	16	0
number of towers with 90% of energy	1	8	8	0
seed pT	1	16	16	0
area	1	16	16	0
algorithm ID	1	16	0	16
unclustered energy	10	32	0	320
for bc-tagged jets tagger mask (allows multiple algorithms)	1	8	8	0

	pointer to lepton used in tag	2	8	16	0
	pointer to tagged vertex	1	8	8	0
	quality of tag	1	16	16	0
	impact parameter discriminant	1	16	16	0
	number of tracks used for discriminant	1	8	8	0
	likelihood probability (b/c/q/gluon)	4	16	64	0
	overall quality (use all information)	1	8	8	0
total(bits)				583	336
total(bytes)				73	42

The jet momentum is determined using the primary vertex (see global object) and the energy in the calorimeter cluster.

The association of tracks and jets cannot be determined from the thumbnail, because it is based on ΔR with individual cells. It should be studied whether 10 track pointers are adequate for all jet flavors (gluon, light and heavy quarks).

Uncertainties of the jet momentum depend on jet energy scale corrections and can be calculated from the thumbnail.

The b/c tag information will be added for all jets for which reco calculates it. This may not include all of the four algorithms. In that case the size estimate has to be revised accordingly.

Missing pT

quantity	values	bits	bits/object	bits/event
missing pT	1	16	16	0
phi	1	16	16	0
sum pT	1	16	16	0
px,y in ECS	2	32	64	0
px,y in ICS	2	32	64	0
px,y in CC	2	32	64	0
px,y in ICN	2	32	64	0
px,y in ECN	2	32	64	0
EM ring	1	624	624	0
FH ring	1	624	624	0
CH ring	1	624	624	0
NADA				
eta/phi	2	8	0	18
layer	1	5	0	6
energy	1	16	0	18
probability	1	8	0	9
total(bits)			2240	50
total(bytes)			280	6

Questions: Do we need to save any errors for the momentum vector?

NADA is the hot cell finding algorithm. We assume 1.1 hot cells/event. This was estimated based on ttbar events (MC), high pT jets (MC), and a sample of very high pT jets from Run 1.

Tracks[ix]

quantity	values	bits	bits/object
track parameters	5	16	80
error matrix	9	16	144
chisquared	1	8	8
CPS/FPS energy	1	8	8
pointer to preshower	1	8	1
CFT pulse height	1	8	8
SMT pulse height	1	8	8
best vertex	1	8	8
hit mask	1	32	32
esc code bit mask	16	1	16
total(bits)			313
total(bytes)			39

Questions: Is the number of degrees of freedom for the chi-squared determined?

Preshower clusters will only be saved for isolated tracks.

We require that the number of 2d clusters in CFT (max=16) and SMT (max=12) hits returned by the D0tracks interface and used by the vertex finding algorithm can be determined from the hit mask stored with every track.

Vertices[x]

quantity	values	bits	bits/object
vertex position	3	32	96
error matrix	6	16	96
chisquared	1	32	32
degrees of freedom	1	16	16
number of tracks	1	16	16
track mask	48	1	48
total(bits)			304
total(bytes)			38

Questions: Is the number of degrees of freedom for the chi-squared determined by the number of tracks?

At this time, the secondary vertex-tagging algorithm does not use the output of the standard vertex finding algorithm but finds its own vertices. It should be verified that all information needed by the secondary tagging algorithm is present on the thumbnail.

Preshower Clusters[xi]

quantity	values	bits	bits/object
energy	1	16	16
phi	1	32	32
phi error	1	16	16

z/r	1	32	32
z/r error	1	16	16
3-d match quality	1	8	8
fraction of energy in layers	3	16	48
profile	4	16	64
width	4	16	64
CPS/FPS flag	1	1	1
total(bits)			297
total(bytes)			37

Only preshower clusters linked to electrons, taus, and isolated tracks will be saved.

Appendix

Other comments received but not adopted

General comments

For each run, include a block that has run-specific information. This presents technical difficulties when selecting events from a file or when combining files. We would instead prefer using a small random access database accessed by the analysis framework.
Make cuts (pT thresholds, isolation) to control number of objects only if absolutely necessary. For example, we probably have to require some isolation for photons, but we would like to keep all tracks

Global

Include global topological quantities, such as H_T, S_T, aplanarity, sphericity. These can be calculated from information in thumbnail. Only include quantities which cannot be recalculated.

Electromagnetic clusters

Add cell level information. Assuming 100 cells per cluster and a 16-bit address and a 16-bit data word per cell, and dropping all other layer information, this would increase the electromagnetic cluster object by 3160 bytes/cluster.
Add track match significance. Contained in global fit chi-squared and can be recalculated.
Add projected vertex from layer centroids. This should be calculated from the stored information at analysis time.

Muons

Include pulse height in muon system and ∫Bdl. There is no useful dE/dx information from the muon system. ∫Bdl has been used only for acceptance cuts. The status word includes a bit for "low ∫Bdl". The muon ID group thinks this will be sufficient and there is no need to keep the value in a separate word.

Taus

Include cell level information. This is too big and beyond the scope of the thumbnail.

Preshower clusters

Add strip information for all 3 or 4 layers. Assuming 11 strips hit per layer for each cluster and three layers, this would increase the size of the preshower object by about 110 bytes/cluster. This includes the savings from dropping the energy fraction, shape and width for each layer.

---

[i] Members of the Data Tier Committee: U. Heintz, J. Hobbs, B. Klima, J. Krane, L. Lueking, S. Protopopescu, M. Strovink, A. Zylberstejn
[ii] Suyong Choi – private communication
[iii] provided by G. Blazey for the trigger group
[iv] presented by S. Protopopescu
[v] presented by D. Zieminska
[vi] presented by Q. Z. Li
[vii] presented by R. Hirosky, private communication from L. Duflot
[viii] provided by R. van Kooten for the b-ID group
[ix] proposed by S. Protopopescu and J. Hobbs
[x] presented by G. Lima
[xi] as suggested by Phil Baringer