1 Algorithm Physics Performance

The physics choices driving the fitting algorithm selection have been described extensively elsewhere[1]. The major points are recapped here. A $\chi^2$ fit is performed in the $r\phi$ plane. The fit function is a linearized form of the equation of a circle expressed in terms of track curvature $\kappa$, impact parameter $b$ and direction at the point of closest approach of the track to the origin $\phi_0$. The function, in polar coordinates $(r, \phi)$, is

$$\phi(r) = b/r + \kappa r + \phi_0.$$ (1)

The fit will return $b, \kappa, \phi_0$ in a coordinate system with the origin at the center of the SMT/CFT assembly. The beam spot may not coincide with this origin. The fit parameters $b$ and $\phi_0$ will have to be corrected to reflect the true beam position. A first pass at this correction has been derived and appears straightforward\(^2\).

The major complexity in the fitting is selecting which hits belong to the track. The average number of hits in a layer of silicon in a given road is significantly more than one. A number of different hit selection algorithms have been tried[1], and most seem to have similar performance. The most significant variation is whether all four SMT layers are required to have hits, or if tracks having only three layers are allowed. Permitting three layers increases the acceptance by roughly 15\%. The final algorithm for hit selection thus uses either four-layer or three-layer candidate tracks, and within each layer, the hit closest to the circular trajectory defined by the two CFT points (from Level 1) and the origin is used in the fit. In addition the hits must either all be in the same barrel segment or in adjacent segments in a logically consistent manner. If a four-layer fit is possible, it is made. If the resulting $\chi^2$ is unacceptably large, the fit is repeated, this time without the hit having the largest contribution to the original four-layer $\chi^2$. This combination of hit selection and two-pass fitting is called the “Static Road/3a” algorithm.

2 Algorithm, Implementation

The fitting algorithm has three phases:

- hit selection
- parameter calculation, and
- $\chi^2$ calculation

This section begins with a brief description of each phase. The next portion details the approximations made to arrive at an algorithm which executes in a reasonable time. The final portion gives the formats and sizes of the major tables (arrays) used in the calculation.

The fit uses at most one hit from each of the four SMT layers. Most of the time, however, some layers have more than one hit in the road. The hits used in the fit are those which

---

\(^1\) The magnitude is the perpendicular distance between the track and origin.

\(^2\) The calculation relies on the approximations $b\kappa << 1$ and $r_V \kappa << 1$. Here $r_V$ is the transverse distance between the beam position and the D0 origin.
Table 1: Fitting algorithm execution time per road in various processors.

<table>
<thead>
<tr>
<th>Processor</th>
<th>Algorithm Type</th>
<th>Execution time (μs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 fit</td>
</tr>
<tr>
<td>Altera 10k130E</td>
<td>float, no LUT</td>
<td>36</td>
</tr>
<tr>
<td>Altera 10k130E</td>
<td>float, w/LUT</td>
<td>18</td>
</tr>
<tr>
<td>TI6701@167 MHz</td>
<td>float, no LUT</td>
<td>17</td>
</tr>
<tr>
<td>TI6701@167 MHz</td>
<td>float, w/LUT</td>
<td>10</td>
</tr>
<tr>
<td>TI6202/3@300 MHz</td>
<td>integer, w/LUT</td>
<td>7.4</td>
</tr>
</tbody>
</table>

are closest to the center of the road. The center is defined as the circle passing through the origin and the centers of the A- and H-layer CTT fibers in the L1CTT track.

The parameter calculation[2] can be reduced to solving the matrix equation\(^2\)

\[
\bar{p} = M \cdot \delta \phi
\]

in which \(\bar{p}\) is the 3-element vector of the fit parameters, \(M\) is a \(3 \times (N_{hit} - 1)\) matrix depending only on the hit radii and resolutions and \(\delta \phi\) is an \(N_{hit} - 1\) element vector containing hit \(\phi\) residuals. The residuals are defined by subtracting the \(\phi\) for one of the hits from all other hits. Thus, one of the hits is guaranteed to have \(\phi = 0\), and the others are nearly zero. The final \(\phi_0\) is obtained by adding back the offset subtracted at the beginning. This method permits a significant reduction in the number of bits required for angles when performing the fit.

The \(\chi^2\) is evaluated using the form

\[
\chi^2 = \sum_{i=1}^{N_{hits}} c_i \left( b - \kappa r_i^2 + (\phi_0 - \phi_i) r_i \right)^2.
\]

Here \(c_i\) is the inverse of the square of the hit resolution in length units.

2.1 Execution Time

An exhaustive number of processing options have been considered, including standard D0 DEC/Compaq alpha processors, fully custom programmable logic devices and digital signal processors. Both floating point and integer versions of the calculation have been tested. Table 1 gives the per track fit times for a variety of processors using the “Static Road/3a” algorithm. The final processor choice, based on execution time, cost, processor density and programming simplicity, is the TI 320C6202/6203 family of 16-bit integer DSP’s.\(^4\) An integer form of the algorithm with matrices stored in a large look up table (LUT) will be used.

\(^3\)The major change in the fitting algorithm over the past six months was the recasting of the original 3x3 form of the matrix \(M\) to the current form. This greatly improved the numerical stability, allowing the precision of the individual entries to be substantially reduced.

\(^4\)The difference between the two models, the '02 and the '03, is the internal memory size. The two are pin compatible, so the final choice between these two will depend on availability and cost.
We expect 2.3 CTT tracks\textsuperscript{5} per 60° sector on average, and six CTT tracks in the highest occupancy sector (again, on average), and we have roughly 50 $\mu$s per event for all fitting.\textsuperscript{6} Thus, a single processor is insufficient for the task. We will build dedicated processor cards each of which processes all tracks in a 30° $\phi$ sector. Each card will have eight DSP’s. If all tracks in a given event require two iterations but assuming no more than one track per processor, it takes 16 $\mu$s to process one event. This time includes a contribution of 5.8 $\mu$s from worst case bus contention and assumes the maximum allowed hit counts in each road. When more than eight tracks are present, the additional fits will be performed as individual processors become available after completing their first fit. The hardware is described in more detail in the following sections.

2.2 Implementation

The best processing times come from the C6202/03 family of DSP’s running an integer form of the algorithm. The integer form is used because floating point arithmetic is emulated in software in the C62 family resulting in unacceptably long execution time if a floating point form is used. Additionally, were the highest precision integer arithmetic used, the dynamic range would be insufficient to calculate the matrix $M$ on-the-fly, so the matrices will be computed offline and stored in a look up table. The DSP naturally performs 16 bit integer multiplies and 32 bit sums. Given this, the algorithm is designed to use matrix elements and hit $\phi$ positions which require only 16 bits. In addition, we currently assume that no more than 32 hits are allowed per road. Monte Carlo studies indicate this is a reasonable upper bound.

Figure 1 gives a comparison between the fit results obtained using a full floating-point calculation and the results from an integerized form. For the integer calculation the matrices were initially calculated in double precision and then converted to the 16 bit integers used in the calculation. The distributions shown are the differences between the integer-based result and the full floating-point result. The data are tracks in the SMT acceptance from 2500 $WH \rightarrow q\overline{q}b\overline{b}$ events. This version of the integer algorithm was designed to produce answers with 1 $\mu$m precision for impact parameter, 1 mrad precision in $\phi_0$, and the curvature result in 256 even bins in the range $-0.003 \leq \kappa \leq 0.003$ cm$^{-1}$. The results for $b$ and $\phi_0$ are substantially more precise than physics demands: the beam spot alone contributes 30$\mu$m to the impact parameter resolutionm the high-$p_T$ track resolution is 15 $\mu$m, and the standard $\phi$ resolution in the trigger is 8 bits or roughly 24.5 mrad. The curvature binning was chosen to match the target output format.\textsuperscript{7}

The previous figures showed the precision of an integer calculation. For those plots the coefficient matrices $M$ were calculated using the true hit radii and then converted to integer format for the calculation. If all possible matrices were computed in this manner and stored in a matrix look up table (MLUT), the table would be far too large. Each road requires a set of matrices because different assigned hit patterns result in numerically different matrices.\textsuperscript{8}

\textsuperscript{5}Based on luminosity weighted QCD events.
\textsuperscript{6}See the results from the queueing simulation elsewhere.
\textsuperscript{7}It could easily be made more precise if needed, for example, in the $\chi^2$ calculation
\textsuperscript{8}For example, the matrix for a four-layer track is different than that with the outer layer not used. Similarly, barrel-to-barrel variation in assembly precision my result in different matrices for tracks in different
Figure 1: The difference in track parameter between the integer and double precision calculations. The resulting RMS values are significantly smaller than what is needed for the trigger. The lower right hand plot indicates whether the hit assignment in the integer algorithms is different from (zero bin) or the same as (one bin) if it were determined using a floating point calculation.
Figure 2: The track parameter resolutions calculated using a matrix look up table having 160 separate $\phi$ slices. The tracks used here were generated with a simple simulation having the correct CTT and SMT geometries but no multiple scattering or false hits. Tracks in the ranges $-0.2 \text{ cm} \leq b \leq 0.2 \text{ cm}$ and $p_T > 1.5 \text{ GeV}$ were generated. For the upper right hand plot only tracks satisfying $p_T > 3 \text{ GeV}$ and $|b| < 0.1 \text{ cm}$ were used. See the text for a discussion of the distribution widths.

If a complete set of matrices were computed for each road, the MLUT for a 30° sector would be 0.5 Gb, clearly too large to be practical. Matrix sets will instead be (pre)computed with one matrix for a range of roads. The final definition of these “super roads” has not yet been determined, but figure 2 show the reconstructed impact parameter distribution for a look up table consisting of 160 $\phi$ super roads, indexed only by H-layer fiber. Within a given super road, all hits in a given sublayer are assumed to come from the same radius. This approximation clearly breaks down for sufficiently wide roads. One sees that the impact parameter distribution in fig. 2 is somewhat wider than the 15 $\mu$m expected in the absence of multiple scattering. This has occurred for tracks with either $p_T < 3 \text{ GeV}$ or $|b| > 0.1 \text{ cm}$. For both of these cases the additional width is not likely to affect the results. However the effect can be substantially reduced by using the CFT A-layer offset in addition to the H-layer in defining the super road. The detailed final format of the super road indexing will barrels.
Table 2: SMT physical address to fit coordinates look up table format. Three spare bits are reserved for the dE/dx information on an event–by–event basis, and the remaining bit is reserved to augment the layer count in the event that additional SMT layers are added. The radius is an offset to the minimum radius for each ladder.

be determined after trigger rate comparisons have been made.

3 Fitting Hardware

The track fitting card (TFC) will be a custom daughter board for the standard STT motherboard. The inputs to the TFC from the FRC and STC will be received on LVDS links using the standard STT PCMPIP receivers. The block diagram of the TFC is shown in figure 3. An input controller will read the data, reformat it on-the-fly as necessary, and store it in DPM memory for DSP access. It will also act as DSP input bus arbiter. The eight DSP’s on a TFC are arranged as two sets of four. Each set of four has a shared input bus and a separate shared output bus. The DSP expansion bus will be used as the input (and initialization) bus, and the DSP external memory bus will be used as the output and external memory bus. A separate output controller will read the fit results from an output DPM, add headers and trailers, and transmit completed fit results to L2 and L3. Simple communication between the input and output controllers will provide event synchronization checking and DSP scheduling control.

The L1CTT data from the FRC will be read by the input controller first, then the SMT hit data from the STC will follow. Both will be read using the “event” format and buffered in dual port memory (DPM) for reading by the DSP’s.

The L1CTT information will be copied unmodified to holding buffers and simultaneously sent to the DSP dual port memory (DPM) input buffers. The L1CTT format is not that used by the fitting algorithms. A look up table will be used to translate from L1CTT to fit format. It is expected that this information will fit in the internal data memory of the DSP’s, so the translation will be part of the fitting algorithm. If the look up tables require external memory, the conversion will be done by the input controller before loading the road information into the DPM buffer.

The SMT hit positions will also be sent unmodified to the holding buffers and also translated from hardware addresses to fitting coordinates via a look up table. The format of the fit coordinates is given in table 2. The translation will occur during the read cycle, so
Figure 3: Track fit card block diagram.
no additional time is lost. The original two bits of hit dE/dx information from the STC will be preserved, and the hit data will be routed to the DPM buffers after the hit translation.

The DPM buffer and input logic will be constructed so that the CTT and SMT data for a given road are in consecutive memory locations. Prototype logic code has been written and simulated, and the translated hits for each road can be placed in consecutive road-based memory locations on the fly even when consecutive received hits are assigned to different roads. However, the first and last hits to arrive from the STC’s may belong to the same road, so all fitting must wait until the hits have all arrived. Once the hit transmission has finished, each DSP will read road data from the DPM buffer, one DSP after the other. Because the fits for each road are stored in consecutive locations, the reads from buffer to internal DSP memory are efficient. The four DSP’s in one set share a common input bus, so the overall processing time will increase slightly because of bus sharing. An upper bound on time needed to read from the buffers into the four DSP’s sharing a bus is 5.6 μs, including bus arbitration. This is for the case in which all roads have the maximum number of hits present.

The fit algorithm is then run independently in each DSP. Each fit requires one or two accesses to the look up table containing the matrices M. The final size of the look up table is not yet determined. It depends on a number of issues, including assembly precision and the ability to make further φ subdivisions within the 30° sectors. If sufficient assembly precision is achieved, the memory size can be substantially reduced. The hit φ values used in the fitting will be computed to an arbitrary precision using the best alignment available, so the assembly precision enters only via the radii used to calculate the matrices stored in the look up tables. Figure 4 shows impact parameter resolution as a function of radial (mis)alignment. Tracks and hits were generated using a simple Monte Carlo and non-ideal geometry. The hits were converted to channel-like coordinates, and then converted back to fit coordinates using a different (ideal) geometry. The resulting impact parameter resolutions are shown. One sees that radial misalignments of up to 200 μm are tolerable. Given the expected assembly precision, this implies that the barrels at negative global z-coordinate can be treated as a unit, and those at positive global z can be treated as a unit. This substantially reduces the matrix look up table size. Table 3 gives the address format for the matrix look up table for the case in which the negative and positive barrels are treated as independent units. The address requires 14 bits, and as previously, each matrix requires 30 bytes, giving a total memory size of 128k addresses of 32 bits each, or 0.5 Mb.

Further reduction can be had if the 30° sectors can be split into two 15° sectors. In this case, the φ range decreases by two fold and it is also likely that the sublayer in a given road is predetermined by the hit CTT fibers. In this case, the number of bits defining the super road index decreases by one, and the sublayer indices can be dropped. The total memory requirement for the MLUT is then only 16 kb. The 6203 DSP has 512 kb of internal data memory and the matrix look up could fit wholly in internal memory, eliminating external memory access during fitting.9

When processing is finished, the DSP requests the output data bus, and the fit results are placed in a DPM memory area with the address determined by the output controller. The input controller is then notified that the DSP is free, and if roads remain to be processed,

9The 6202 has 128 kb of internal data memory.
Figure 4: Alignment constraints. This figure shows the impact parameter residuals for cases in which two of the four SMT layers are at incorrect radii. The radial shifts are taken to be 100, 200 and 300 μm respectively. This figure is preliminary but exemplifies the method by which tolerances will be derived. The analysis will be extended to understand the all assembly tolerances.
<table>
<thead>
<tr>
<th># of bits</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Super Road ID (160 roads, 30° logical sectors)</td>
</tr>
<tr>
<td>1</td>
<td>Positive or negative z barrel in 4th layer</td>
</tr>
<tr>
<td>2</td>
<td>Barrel transition layer (0, no change)*</td>
</tr>
<tr>
<td>4</td>
<td>Sublayer hit within each layer</td>
</tr>
<tr>
<td>1</td>
<td>Skipping a layer (Y/N)</td>
</tr>
<tr>
<td>2</td>
<td>ID of skipped layer</td>
</tr>
<tr>
<td>3</td>
<td>Matrix element pair (See caption)</td>
</tr>
<tr>
<td>17</td>
<td>TOTAL = 128k addresses for 32-bit words (0.5 Mb)</td>
</tr>
</tbody>
</table>

Table 3: Matrix inverse look up table address format. The look up table is assumed to be generated separately for each 30° sector. If additional layers are to be accommodated, the items with asterisks will have to be expanded by 1 bit. Each matrix has 15 elements of 16 bytes each, or an additional 3 bits of word address. A sector actually covers 360°/8 = 45°. If the mechanical assembly precision is insufficient to treat the negative and positive barrels as single units, an additional 4 bits of address would be needed.

the DSP is notified and begins processing the next road.

All output formatting for the individual track is done as part of the fit. When all tracks in a given sector are finished, the data are read from the buffer by an output controller. A standard header and trailer are added, and the data transmitted to L2 via Cypress hot link. The data format is shown in table 4. For most events, the STT will add 8 bytes of data per road (plus header and trailer). Using the plots from reference [1], the average sector has 16 bytes of fit output, and the highest occupancy sector will have 48 bytes.

Under special conditions signaled by qualifier bits in the original transmission from the FRC, the input hits and road data will also be transmitted. The number of 32-bit words transmitted in this case is $(1 + 1) \times \Sigma_{\text{roads}} N_h + (1 + 3) \times N_\text{roads}$. The first term has two words for each SMT hit. The first is the data received from the STC, and the second is the converted coordinate used by the fitting. The second term has one word for the original L1CTT input and the three words used in the fitting. For these events, the average data size will be roughly 160 bytes per sextant, with a strict per sextant maximum of 11.4 kb.

The output data bus and external matrix look up table share the DSP external memory interface, so matrix reads and data output cannot occur simultaneously. However, each look up and output write are expected to take less than 0.5 μs, to be compared with the overall processing time of roughly 10 μs. It appears that this will not cause problems.

In addition to the STT information with its header/trailer, the original L1CTT information will also be transmitted from the first of the two TFC’s in a sector. Thus, one cable will have both L1CTT and STT header/trailer blocks for each event, and the second will have only the STT information. A number of other options have been considered but lead to nearly doubling the transmitted data size at high luminosity and also require additional hardware complexity. Since the matrix look up tables will be specific to a 30° sector, additional control words at the end of the LUT can be used to flag whether a TFC transmits CTT+STT information or only STT data.
Table 4: STT data block to Level 2 format. Two 32-bit words are used per road. If the beam position correction is made in the TFC, one might use some spare bits to pass the impact parameter significance.

The Level 3 output data for normal events will be a copy of the Level 2 information. For diagnostic events, the complete input to the DSP’s will be added. This information will be placed in the motherboard buffers via the 64-bit PCI interface.

3.1 Power Consumption
The fitting daughterboards will contain eight DSP’s, probably three large pin-count programmable logic devices and up to three significant external memories. The power consumption, based on manufacturers data sheets, is estimated to be 30 W per daughterboard.

3.2 Cost Estimate and schedule
The dominant portions of the cost are the processors, the programmable devices and memories. The cost for these components is roughly evenly divided and the total is $3500. The board manufacturing and assembly is estimated to add another $1000. Thus, the cost for each daughterboards is $4500. The original estimate was $7k/slot.

3.3 Monitoring
Monitoring of the input and output buffers will be provided on the receiver cards. Six CPU states will be defined: idle, input bus request, input data transfer, processing, output bus request and output data transfer. These will be monitored for all CPU’s by the I/O controllers. The states will be latched at fixed time intervals into a single register and read
out when requested. This will allow a stochastic sampling of the CPU states. In addition, monitoring internal to the DSP’s will include

- hits overflow count
- layer overflow count
- empty CTT track count
- intermediate calculation results for diagnostic readout

These will be written to the output DPM buffer upon receipt of appropriate SCL qualifiers (from the FRC). All monitoring information will be accessible from VME.

4 Work Remaining

This section lists issues to be addressed in the immediate future:

- Where will the beam position correction be made, and what is the feedback path from L3?
- alignment
- Initialization
- L2STT to L2CTT header/trailer format. This is a DZero standard
- Recheck all numbers and extend assembly precision constraints with new trigsim as soon as it’s available.

References
