# Inter Hall Open Software submission: Final Report (Slot 1)

## Installation instructions:

Open the terminal. Navigate to the folder plotToTables and install by using the following commands

```
cd .../plotToTables
chmod +x plotsToTables.sh
./plotsToTables.sh
```

***Note:*** A directory having all the files of OpenCV 3.1.0 has been included to avoid manual download from the GitHub repository, effectively working to reduce the time required for installation of the software.

## User Manual:

1. Click on open to select the pdf/jpg/png file.

2. Click on print tables to generate the corresponding tables.

## Description of algorithms used :

- **Identifying the plot area:** In order to identify individual plots in the complete image of the scanned page, a *hough transform* operation is performed to store the horizontal and vertical lines on the page. An individual plot will be characterised by 4 vertices corresponding to the intersection points of it's horizontal and vertical axes ; also, the corresponding axes labels and markers will be present to the left of the y axes and to the bottom of the x axes. Accordingly, pairs of 4 points corresponding to each graph are identified by conditioning on the intersection points and the location of possible text to the left and bottom of the candidate vertical and horizontal axes. Reading of text is done by integrating the functionality of Tesseract, a commonly available Optical Character Recognition Library within the code. The information related to the required co-ordinates of each of the plots and the scale information across both axes acquired at this stage.

- **Reading axes labels:** From the previous steps, the information about the spatial location of the origin of the plot is available to us. A single *dilation* is now performed to *smoothen* the labels and markers. A scan is performed along the x direction (to left of the origin) to identify the width of markers along the y axis, based on which a *bounding rectangle* is drawn to capture the information of the y scale(fed to Tesseract), following which the axis label is also read by a similar method involving scanning to the left of the y marker rectangle boundary. The same logic is used to acquire the x axis data.

- **Extracting the image window:** With the co-ordinates obtained from the *hough line* based axis detection technique, a window spanning the area of interest (graph area) is extracted.

This area corresponds to one of the graphs in the image.

- **Extracting individual plots from the graph:** For obtaining the data points from each of the plots, it is first necessary to segment out individual plots. This is achieved by a colour based segmentation applied on the graph area.

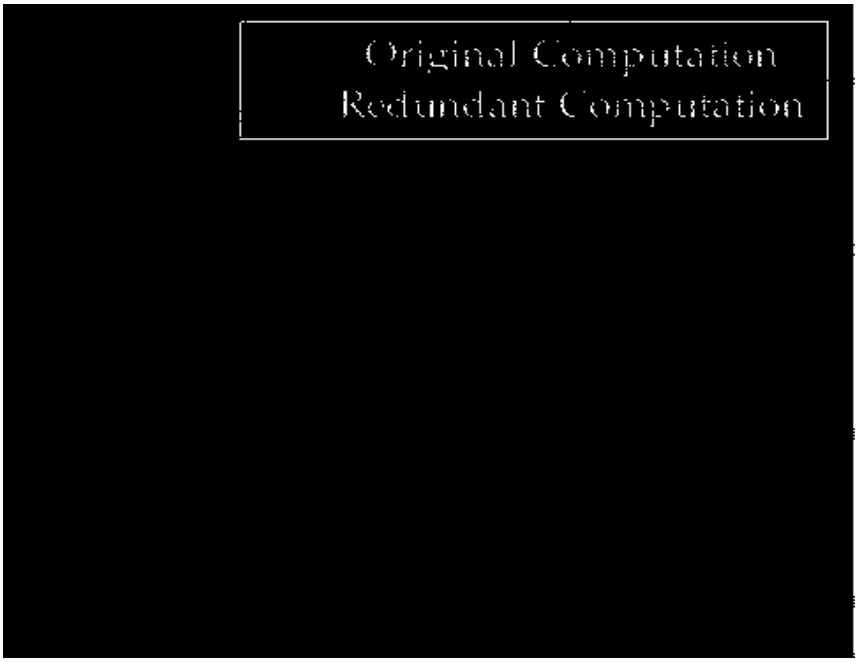



Fig 1 : Mask corresponding to the legend

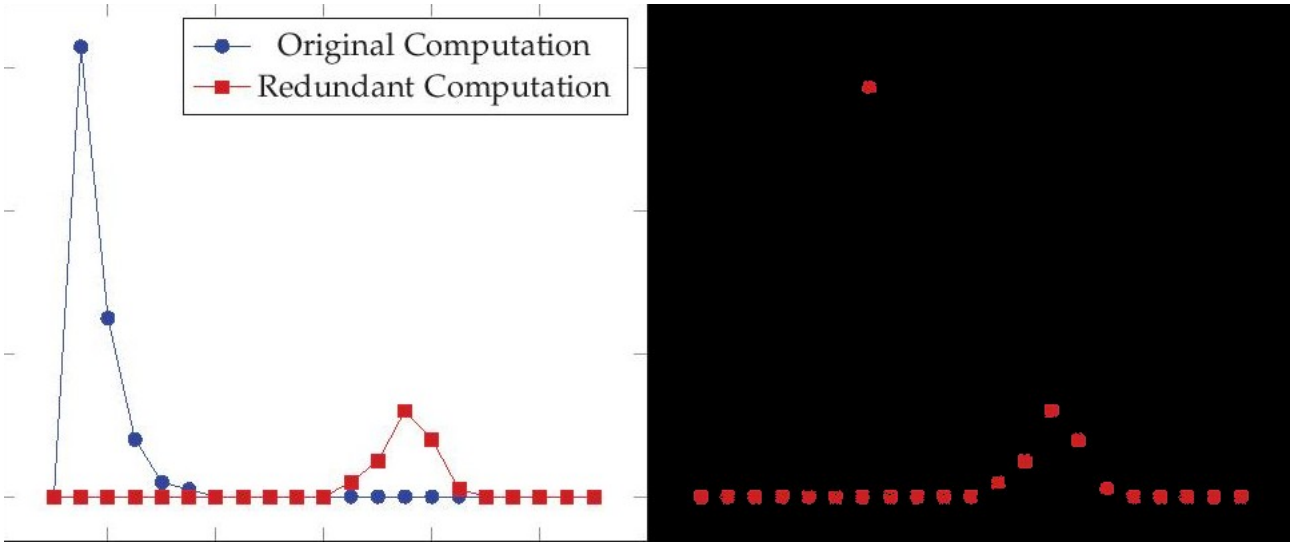



Fig 2: Segmentation result of a single coloured plot

- **Colour based segmentation:**

1. The outer boundaries of the graph belong to black while the background is associated with a white colour. Assuming a colour bin to be of plus or minus 40 pixel range, initially a *mask* is created out of the segmented black region, which is expected to contain the text, boundaries and possible plots corresponding to a black legend. A temporary mask is prepared from the

output of this step. In order to construct a bounding rectangle for the legend, we scan the thresholded and binarised version of the mask across both the x and y directions, effectively detecting the locations of the co-ordinates of the span of the legend text or border using which the legend area is then *segmented* out.

2. For the purpose of segmenting out each of the plots, it is required that the colour of the plot be identified. This is done by finding the frequency of occurence of all the colours present in the image and separating out the top hundred colour candidates based on their frequency, followed by an extraction of the individual plot colours by integrating the results of these candidates based on a *threshold* applied on each of the individual channels (*bucket* of colour values which correspond to a single plot colour) and segmentation into separate images. During segmentation, an elimination of all gray candidates along with the black and white pixels must also be performed in order to avoid false positives.

3. After obtaining the individual segmented plot lines, we attempt to connect the missing regions of the plots by performing an opening operation. Using the information obtained from the scale and the location of zero co-ordinates of the axes for a plot, we construct the data table by perfoming a sampling the on the x axis in incremental steps of one tenth the value of the x scale. Sampling is perfomed by scanning across the y line at a particular x value and using y scale information to determine the corresponding y value. The points from the legend area have to be excluded in this mapping.

4. In order to tackle the case of missing values in the plot, we perform a bilinear interpolation using the stored values from either side of the missing point. In case of missing values at the start and end of the plot, the corresponding values obtained from the right and left are taken respectively to fill the table.

**Marking plot labels:** From the detected legend area, using the spatial information of the text labels, we assign labels to each of the plots based on the colour threshold values obtained from the colour based segmentation performed on the plot area.

## Extras:

**Input form conversion:** The input given by the user would be in the form of a portable document format (PDF) which is converted into an image by employing pyPdf [1] and Wand [2] (an Image Magick binding available in python). The converted image can now be read and processed by the use of Python OpenCV [3].
**Text Reader:** Text is read from the image by employing Tesseract[4], a freely downloadable Character Recognition Library.

**Limitations:**

1. The accuracy of the text reader (Tesseract) is highly dependent on the resolution of the image supplied and gives erroneous results for poorly rendered images.

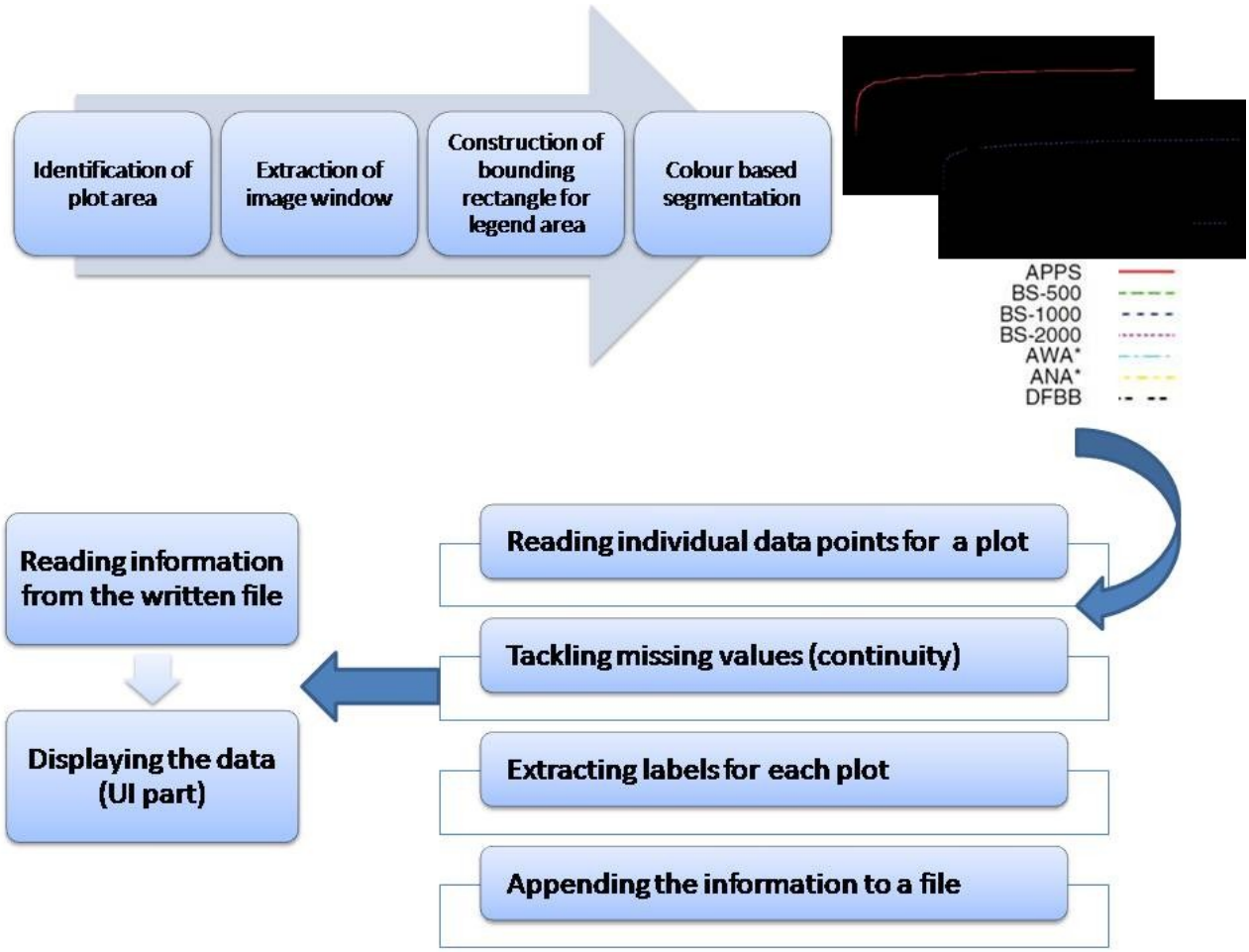
2. Plots having pixel values very close to white (255,255,255) or black (0,0,0) may not be segmented properly resulting in too many missing values giving rise to problems while extracting plot values.

3. In cases where the individual plots in a graph are not properly separated in the beginning and at the end of the x range, the interpolation is not properly performed resulting in missing values in those ranges.

## References:


[1] https://pypi.python.org/pypi/pyPdf/1.13
[2] https://pypi.python.org/pypi/Wand
[3] http://docs.opencv.org/3.0-beta/index.html
[4] https://github.com/tesseract-ocr

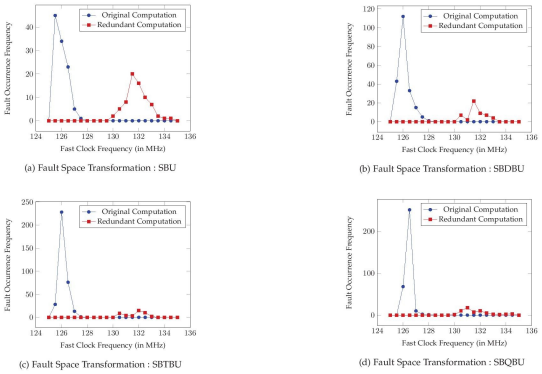

## Architectural diagram :



# Test Cases:

1. The pdf contains 6 images, each having 2 individual plots having individual data point markers for each of the plots.The legend is enclosed in a bounding box. The input image and the results generated by running the pipeline are indicated below.

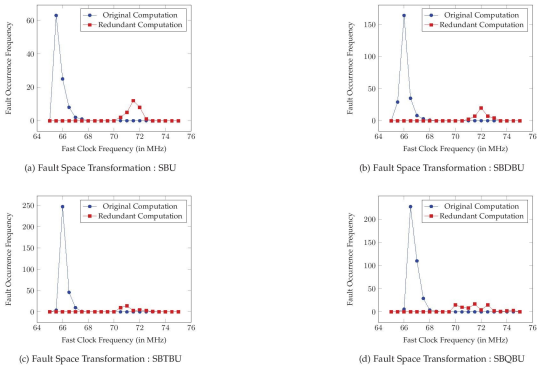





(a) Fault Space Transformation : SBU (b) Fault Space Transformation : SBDBU

(c) Fault Space Transformation : SBTBU (d) Fault Space Transformation : SBQBU

Fig. 9: Effect of Fault Space Transformation on the Time Redundancy Countermeasure

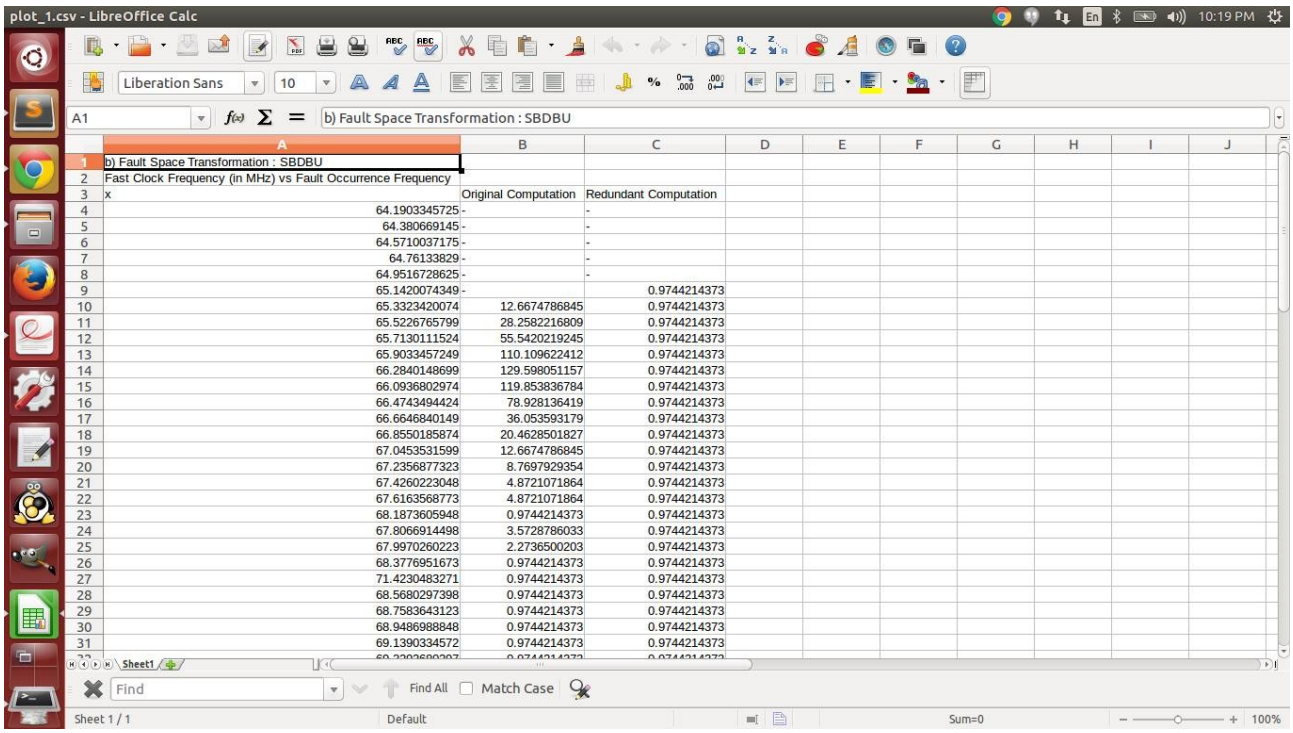



(a) Fault Space Transformation : SBU (b) Fault Space Transformation : SBDBU

(c) Fault Space Transformation : SBTBU (d) Fault Space Transformation : SBQBU

Fig. 10: Effect of Fault Space Transformation on the Hardware Redundancy Countermeasure

Fig 3 : Input pdf file for case 1



plot_1.csv - LibreOffice Calc

b) Fault Space Transformation : SBDBU

| | A | B | C |
|---|---|---|---|
| 1 | b) Fault Space Transformation : SBDBU | | |
| 2 | Fast Clock Frequency (in MHz) vs Fault Occurrence Frequency | | |
| 3 | x | Original Computation | Redundant Computation |
| 4 | 64.1903345725 | - | - |
| 5 | 64.380669145 | - | - |
| 6 | 64.5710037175 | - | - |
| 7 | 64.76133829 | - | - |
| 8 | 64.9516728625 | - | - |
| 9 | 65.1420074349 | - | 0.9744214373 |
| 10 | 65.3323420074 | 12.6674786845 | 0.9744214373 |
| 11 | 65.5226765799 | 28.2582216809 | 0.9744214373 |
| 12 | 65.7130111524 | 55.5420219245 | 0.9744214373 |
| 13 | 65.9033457249 | 110.109622412 | 0.9744214373 |
| 14 | 66.2840148699 | 129.598051157 | 0.9744214373 |
| 15 | 66.0936802974 | 119.853836784 | 0.9744214373 |
| 16 | 66.4743494424 | 78.928136419 | 0.9744214373 |
| 17 | 66.6646840149 | 36.053593179 | 0.9744214373 |
| 18 | 66.8550185874 | 20.4628501827 | 0.9744214373 |
| 19 | 67.0453531599 | 12.6674786845 | 0.9744214373 |
| 20 | 67.2356877323 | 8.7697929354 | 0.9744214373 |
| 21 | 67.4260223048 | 4.8721071864 | 0.9744214373 |
| 22 | 67.6163568773 | 4.8721071864 | 0.9744214373 |
| 23 | 68.1873605948 | 0.9744214373 | 0.9744214373 |
| 24 | 67.8066914498 | 3.5728786033 | 0.9744214373 |
| 25 | 67.9970260223 | 2.2736500203 | 0.9744214373 |
| 26 | 68.3776951673 | 0.9744214373 | 0.9744214373 |
| 27 | 71.4230483271 | 0.9744214373 | 0.9744214373 |
| 28 | 68.5680297398 | 0.9744214373 | 0.9744214373 |
| 29 | 68.7583643123 | 0.9744214373 | 0.9744214373 |
| 30 | 68.9486988848 | 0.9744214373 | 0.9744214373 |
| 31 | 69.1390334572 | 0.9744214373 | 0.9744214373 |

Fig 4: Generated output for case 1

2. The pdf document contains a single graph consisting of 6 plots. However, the legend indicates 7 plots, one of which is of black colour. Accordingly, no data points are generated for that label. The input file and the generated output are indicated below.

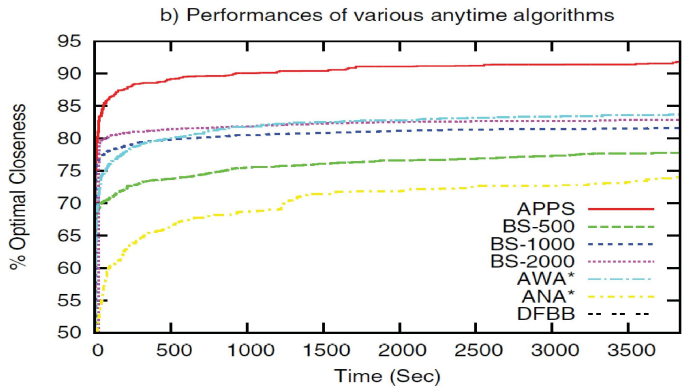
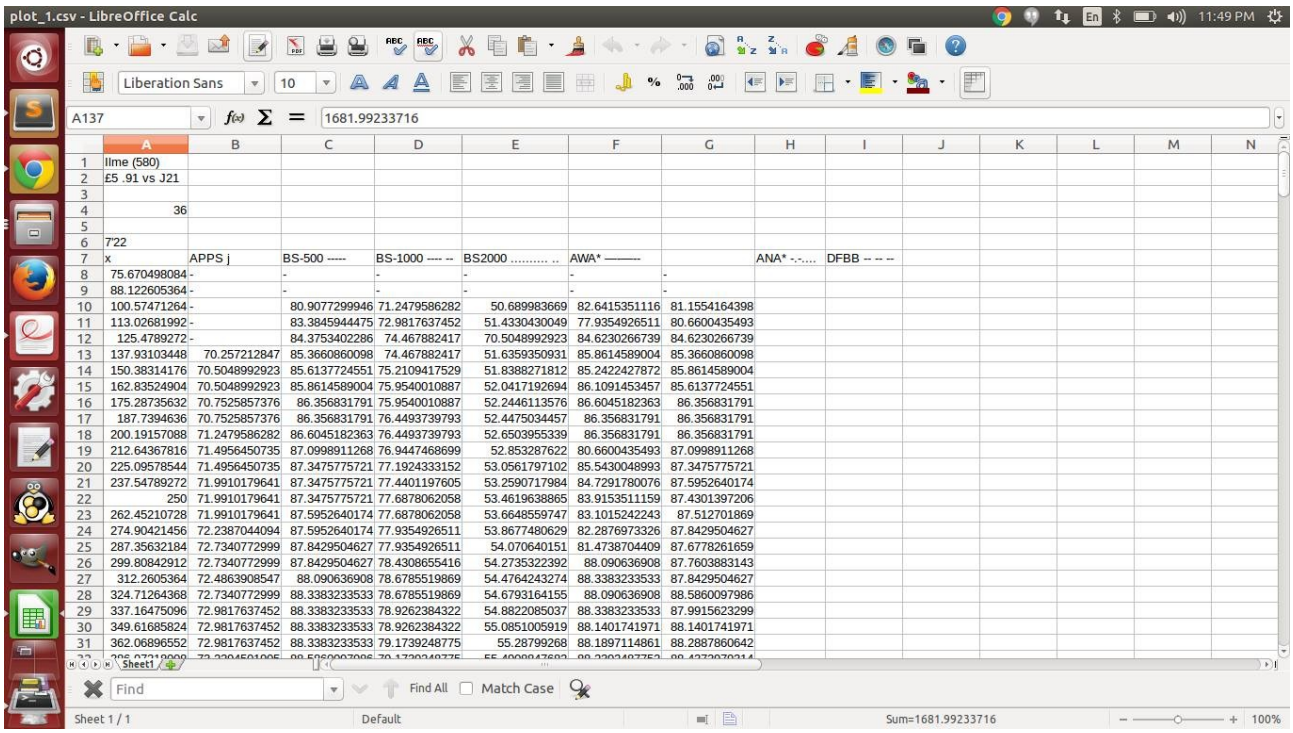



Fig 5: Input pdf file for case 2




| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Ilme (580) | | | | | | | | | | | | | |
| 2 | £5 .91 vs J21 | | | | | | | | | | | | | |
| 3 | | | | | | | | | | | | | | |
| 4 | 36 | | | | | | | | | | | | | |
| 5 | | | | | | | | | | | | | | |
| 6 | 7'22 | | | | | | | | | | | | | |
| 7 | x | APPS j | BS-500 ----- | BS-1000 ---- -- | BS2000 ..... .. | AWA* ------- | | ANA* -.-.... | DFBB -- -- -- | | | | | |
| 8 | 75.670498084 | - | - | - | - | - | - | | | | | | | |
| 9 | 88.122605364 | - | - | - | - | - | - | | | | | | | |
| 10 | 100.57471264 | - | 80.9077299946 | 71.2479586282 | 50.689983669 | 82.6415351116 | 81.1554164398 | | | | | | | |
| 11 | 113.02681992 | - | 83.3845944475 | 72.9817637452 | 51.4330430049 | 77.9354926511 | 80.6600435493 | | | | | | | |
| 12 | 125.4789272 | - | 84.3753402286 | 74.467882417 | 70.5048992923 | 84.6230266739 | 84.6230266739 | | | | | | | |
| 13 | 137.93103448 | 70.257212847 | 85.3660860098 | 74.467882417 | 51.6359350931 | 85.8614589004 | 85.3660860098 | | | | | | | |
| 14 | 150.38314176 | 70.5048992923 | 85.6137724551 | 75.2109417529 | 51.8388271812 | 85.2422427872 | 85.8614589004 | | | | | | | |
| 15 | 162.83524904 | 70.5048992923 | 85.8614589004 | 75.9540010887 | 52.0417192694 | 86.1091453457 | 85.6137724551 | | | | | | | |
| 16 | 175.28735632 | 70.7525857376 | 86.356831791 | 75.9540010887 | 52.2446113576 | 86.6045182363 | 86.356831791 | | | | | | | |
| 17 | 187.7394636 | 70.7525857376 | 86.356831791 | 76.4493739793 | 52.4475034457 | 86.356831791 | 86.356831791 | | | | | | | |
| 18 | 200.19157088 | 71.2479586282 | 86.6045182363 | 76.4493739793 | 52.6503955339 | 86.356831791 | 86.356831791 | | | | | | | |
| 19 | 212.64367816 | 71.4956450735 | 87.0998911268 | 76.9447468699 | 52.853287622 | 80.6600435493 | 87.0998911268 | | | | | | | |
| 20 | 225.09578544 | 71.4956450735 | 87.3475775721 | 77.1924333152 | 53.0561797102 | 85.5430048993 | 87.3475775721 | | | | | | | |
| 21 | 237.54789272 | 71.9910179641 | 87.3475775721 | 77.4401197605 | 53.2590717984 | 84.7291780076 | 87.5952640174 | | | | | | | |
| 22 | 250 | 71.9910179641 | 87.3475775721 | 77.6878062058 | 53.4619638865 | 83.9153511159 | 87.4301397206 | | | | | | | |
| 23 | 262.45210728 | 71.9910179641 | 87.5952640174 | 77.6878062058 | 53.6648559747 | 83.1015242243 | 87.512701869 | | | | | | | |
| 24 | 274.90421456 | 72.2387044094 | 87.5952640174 | 77.9354926511 | 53.8677480629 | 82.2876973326 | 87.8429504627 | | | | | | | |
| 25 | 287.35632184 | 72.7340772999 | 87.8429504627 | 77.9354926511 | 54.070640151 | 81.4738704409 | 87.6778261659 | | | | | | | |
| 26 | 299.80842912 | 72.7340772999 | 87.8429504627 | 78.4308655416 | 54.2735322392 | 88.090636908 | 87.7603883143 | | | | | | | |
| 27 | 312.2605364 | 72.4863908547 | 88.090636908 | 78.6785519869 | 54.4764243274 | 88.3383233533 | 87.8429504627 | | | | | | | |
| 28 | 324.71264368 | 72.7340772999 | 88.3383233533 | 78.6785519869 | 54.6793164155 | 88.090636908 | 88.5860097986 | | | | | | | |
| 29 | 337.16475096 | 72.9817637452 | 88.3383233533 | 78.9262384322 | 54.8822085037 | 88.3383233533 | 87.9915623299 | | | | | | | |
| 30 | 349.61685824 | 72.9817637452 | 88.3383233533 | 78.9262384322 | 55.0851005919 | 88.1401741971 | 88.1401741971 | | | | | | | |
| 31 | 362.06896552 | 72.9817637452 | 88.3383233533 | 79.1739248775 | 55.28799268 | 88.1897114861 | 88.2887860642 | | | | | | | |

Fig 6: output for case 2