Character and String Representation

CS520

Department of Computer Science
University of New Hampshire
CDC 6600

- 6-bit character encodings
- i.e. only 64 characters
- Designers were not too concerned about text processing!

The table is from *Assembly Language Programming for the Control Data 6000 series and the Cyber 70 series* by Grishman.
<table>
<thead>
<tr>
<th>Decimal</th>
<th>Binary</th>
<th>Octal</th>
<th>Hex</th>
<th>ASCII</th>
<th>Decimal</th>
<th>Binary</th>
<th>Octal</th>
<th>Hex</th>
<th>ASCII</th>
<th>Decimal</th>
<th>Binary</th>
<th>Octal</th>
<th>Hex</th>
<th>ASCII</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>00000000</td>
<td>000 00</td>
<td>NUL</td>
<td></td>
<td>32</td>
<td>00100000</td>
<td>040 20</td>
<td>SP</td>
<td></td>
<td>64</td>
<td>01000000</td>
<td>100 40</td>
<td>@</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>00000001</td>
<td>001 01</td>
<td>SOH</td>
<td>!</td>
<td>33</td>
<td>00100001</td>
<td>041 21</td>
<td>!</td>
<td></td>
<td>65</td>
<td>01000001</td>
<td>101 41</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>00000010</td>
<td>002 02</td>
<td>STX</td>
<td>&quot;</td>
<td>34</td>
<td>00100010</td>
<td>042 22</td>
<td>&quot;</td>
<td></td>
<td>66</td>
<td>01000010</td>
<td>102 42</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>00000011</td>
<td>003 03</td>
<td>ETX</td>
<td>#</td>
<td>35</td>
<td>00100011</td>
<td>043 23</td>
<td>#</td>
<td></td>
<td>67</td>
<td>01000011</td>
<td>103 43</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>00000100</td>
<td>004 04</td>
<td>EOT</td>
<td>$</td>
<td>36</td>
<td>00100100</td>
<td>044 24</td>
<td>$</td>
<td></td>
<td>68</td>
<td>01001000</td>
<td>104 44</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>00000101</td>
<td>005 05</td>
<td>ENQ</td>
<td>&amp;</td>
<td>37</td>
<td>00100101</td>
<td>045 25</td>
<td>&amp;</td>
<td></td>
<td>69</td>
<td>01001010</td>
<td>105 45</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>00000110</td>
<td>006 06</td>
<td>ACK</td>
<td>'</td>
<td>38</td>
<td>00100110</td>
<td>046 26</td>
<td>'</td>
<td></td>
<td>70</td>
<td>01001100</td>
<td>106 46</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>00000111</td>
<td>007 07</td>
<td>BEL</td>
<td>(</td>
<td>39</td>
<td>00100111</td>
<td>047 27</td>
<td>(</td>
<td></td>
<td>71</td>
<td>01001110</td>
<td>107 47</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>00001000</td>
<td>010 08</td>
<td>BS</td>
<td>)</td>
<td>40</td>
<td>00101000</td>
<td>050 28</td>
<td>)</td>
<td></td>
<td>72</td>
<td>01010000</td>
<td>110 48</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>00001001</td>
<td>011 09</td>
<td>HT</td>
<td>)</td>
<td>41</td>
<td>00101001</td>
<td>051 29</td>
<td>)</td>
<td></td>
<td>73</td>
<td>01010010</td>
<td>111 49</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td></td>
<td>00001010</td>
<td>012 10</td>
<td>LF</td>
<td>2A</td>
<td>42</td>
<td>00101010</td>
<td>052 30</td>
<td>2A</td>
<td></td>
<td>74</td>
<td>01010100</td>
<td>112 50</td>
<td>J</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>00001011</td>
<td>013 11</td>
<td>VT</td>
<td>2B</td>
<td>43</td>
<td>00101011</td>
<td>053 31</td>
<td>2B</td>
<td></td>
<td>75</td>
<td>01010110</td>
<td>113 51</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>00001100</td>
<td>014 12</td>
<td>FF</td>
<td>2C</td>
<td>44</td>
<td>00101100</td>
<td>054 32</td>
<td>2C</td>
<td></td>
<td>76</td>
<td>01011100</td>
<td>114 52</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>00001101</td>
<td>015 13</td>
<td>CR</td>
<td>2D</td>
<td>45</td>
<td>00110110</td>
<td>055 33</td>
<td>2D</td>
<td></td>
<td>77</td>
<td>01100110</td>
<td>115 53</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>00001110</td>
<td>016 14</td>
<td>SO</td>
<td>2E</td>
<td>46</td>
<td>00111000</td>
<td>056 34</td>
<td>2E</td>
<td></td>
<td>78</td>
<td>01100111</td>
<td>116 54</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>00001111</td>
<td>017 15</td>
<td>SI</td>
<td>2F</td>
<td>47</td>
<td>00111100</td>
<td>057 35</td>
<td>2F</td>
<td></td>
<td>79</td>
<td>01101110</td>
<td>117 55</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td></td>
<td>00010000</td>
<td>020 16</td>
<td>DLE</td>
<td>30</td>
<td>48</td>
<td>01000000</td>
<td>060 36</td>
<td>30</td>
<td></td>
<td>80</td>
<td>01010000</td>
<td>120 50</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>00010001</td>
<td>021 17</td>
<td>DC1</td>
<td>31</td>
<td>49</td>
<td>01010001</td>
<td>061 37</td>
<td>31</td>
<td></td>
<td>81</td>
<td>01010010</td>
<td>121 51</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>00010010</td>
<td>022 18</td>
<td>DC2</td>
<td>32</td>
<td>50</td>
<td>01010010</td>
<td>062 38</td>
<td>32</td>
<td></td>
<td>82</td>
<td>01010100</td>
<td>122 52</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>00010100</td>
<td>024 20</td>
<td>DC4</td>
<td>34</td>
<td>52</td>
<td>01010100</td>
<td>064 36</td>
<td>34</td>
<td></td>
<td>84</td>
<td>01100100</td>
<td>124 54</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>00010101</td>
<td>025 21</td>
<td>NAK</td>
<td>35</td>
<td>53</td>
<td>01011000</td>
<td>065 37</td>
<td>35</td>
<td></td>
<td>85</td>
<td>01101010</td>
<td>125 55</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>00010110</td>
<td>026 22</td>
<td>SYN</td>
<td>37</td>
<td>54</td>
<td>01100100</td>
<td>066 38</td>
<td>37</td>
<td></td>
<td>87</td>
<td>01101100</td>
<td>127 57</td>
<td>W</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>00010111</td>
<td>027 23</td>
<td>ETB</td>
<td>39</td>
<td>57</td>
<td>01101000</td>
<td>070 38</td>
<td>39</td>
<td></td>
<td>89</td>
<td>01101100</td>
<td>131 59</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>00011000</td>
<td>031 25</td>
<td>EM</td>
<td>40</td>
<td>58</td>
<td>01110000</td>
<td>072 3A</td>
<td>40</td>
<td></td>
<td>90</td>
<td>01110100</td>
<td>132 5A</td>
<td>Z</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>00011011</td>
<td>033 27</td>
<td>ESC</td>
<td>3B</td>
<td>59</td>
<td>01110100</td>
<td>073 3B</td>
<td>3B</td>
<td></td>
<td>91</td>
<td>01110110</td>
<td>133 5B</td>
<td>[</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>00011100</td>
<td>034 28</td>
<td>FS</td>
<td>3C</td>
<td>60</td>
<td>01111000</td>
<td>074 3C</td>
<td>3C</td>
<td></td>
<td>92</td>
<td>01111010</td>
<td>134 5C</td>
<td>\</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>00011101</td>
<td>035 29</td>
<td>GS</td>
<td>3D</td>
<td>61</td>
<td>01111100</td>
<td>075 3D</td>
<td>3D</td>
<td></td>
<td>93</td>
<td>01111101</td>
<td>135 5D</td>
<td>]</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>00011110</td>
<td>036 30</td>
<td>RS</td>
<td>3E</td>
<td>62</td>
<td>01111101</td>
<td>076 3E</td>
<td>3E</td>
<td></td>
<td>94</td>
<td>01111110</td>
<td>136 5E</td>
<td>^</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>00011111</td>
<td>037 31</td>
<td>US</td>
<td>3F</td>
<td>63</td>
<td>01111110</td>
<td>077 3F</td>
<td>3F</td>
<td></td>
<td>95</td>
<td>01111111</td>
<td>137 5F</td>
<td>_</td>
<td></td>
</tr>
</tbody>
</table>

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/3.0/
C Strings

• Usually implemented as a series of ASCII characters terminated by a null byte (0x00).
• "abc" in memory is:
Unicode

• The space of values is divided into 17 planes.

• Plane 0 is the Basic Multilingual Plane (BMP).
  – Supports nearly all modern languages.
  – Encodings are 0x0000-0xFFFF.

• Planes 1-16 are supplementary planes.
  – Supports historic scripts and special symbols.
  – Encodings are 0x10000-0x10FFFF.

• Planes are divided into blocks.
Unicode and ASCII

- **ASCII is the bottom block in the BMP**, known as the Basic Latin block.
- So ASCII values are embedded “as is” into Unicode.
- i.e. 'a' is 0x61 in ASCII and 0x0061 in Unicode.
Special Encodings

• The Byte-Order Mark (BOM) is used to signal endian-ness.
• Has no other meaning (i.e. usually ignored).
• Encoded as 0xFEFF.
• 0xFFFE is a noncharacter.
  ─ Cannot appear in any exchange of Unicode.
• So file can be started with a BOM; the reader can then know the endian-ness of the file.
• In absence of a BOM, Big endian is assumed.
Other Noncharacters

• There are a total of 66 noncharacters:
  – 0xFFE and 0xFFFF of the BMP
  – 0x1FFE and 0x1FFFF of plane 1
  – 0x2FFE and 0x2FFFF of plane 2
  – etc., up to
  – 0x10FFE and 0x10FFFF of plane 16
  – Also 0xFDD0-0xFDEF of the BMP.
UTF: UCS* Transformation Format

• **UTF-8**
  – Encodes Unicode characters in 1-4 bytes.
  – ASCII gets encoded as 1 byte.
  – Dominant character encoding for the WWW.

• **UTF-16**
  – Encodes BMP characters in 2 bytes
  – Encodes non-BMP characters in 4 bytes.

• **UTF-32**
  – Fixed-sized representation of Unicode.

*Universal Character Set.*
UTF-8

- Take the Unicode character and throw away the leading zero bits.*
- Count the remaining number of bits.
  - 7 bits: 0xxxxxxxx
  - 11 bits: 110xxxxxx 10xxxxxx
  - 16 bits: 1110xxxx 10xxxxxx 10xxxxxx
  - 21 bits: 11110xxx 10xxxxxx 10xxxxxx 10xxxxxx

*Overlong encodings are forbidden. Therefore there is a unique UTF-8 encoding for each Unicode character.
Errors in UTF-8

• Overlong encodings.
• An unexpected continuation byte.
• A start byte not followed by enough continuation bytes.
• A 4-byte sequence starting with 0xF4 that decodes to a value greater than 0x10FFFF.
• A sequence that decodes to a noncharacter.
• A sequence that decodes to a value in range 0xD800-0xDFFF.
UTF-16

• 1 UTF-16 code unit (2 8-bit bytes) for each BMP character.

• 2 UTF-16 code units for each non-BMP character (4 bytes in total).
  – 0x10000 is subtracted from the value, leaving a 20-bit number in the range 0x00000-0xFFFFF.
  – The top 10 bits are added to 0xD800 to give the first code unit, called the *lead surrogate*.
  – The low 10 bits are added to 0xDC00 to give the second code unit, called the *trail surrogate*.
Self-synchronizing

• 10 bits express values in the range 0x000-0x3FF.
• Lead surrogates will be in range 0xD800+0x000 to 0xD800+0x3FF (0xD800-0xDBFF).
• Trail surrogates will be in range 0xDC00+0x000 to 0xDC00+0x3FF (0xDC00-0xDFFF).
• Remember: values 0xD800-0xDFFF are not valid Unicode characters.
• UTF-16 BMP characters can be distinguished from UTF-16 non-BMP characters.
• So you can tell where the Unicode character boundaries are in a UTF-16 stream.
UTF-32

- Simply take the 21-bit Unicode value and add leading zero bits to extend it to 32 bits.
- Byte-order is an issue, like with UTF-16.