Calcolatori Elettronici I

Prof. Roberto Canonico

Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione

A.A. 2019-2020

Codice ASCII

- ASCII (*American Standard Code for Information Interchange*) è un codice standard per la codifica di caratteri definito negli anni '60 dall'*American National Standards Institute* (ANSI)
- Il set di caratteri rappresentabile dal codice ASCII comprende solo i caratteri dell'alfabeto latino standard più una serie di caratteri di controllo (non stampabili) come CarriageReturn (CR), LineFeed (LF), ecc.
- ASCII rappresenta ciascun carattere mediante un numero naturale (*code-point*) tra 0 e 127
- Pur potendo essere rappresentati mediante solo 7 bit, i code-point ASCII sono rappresentati su 8 bit, con il bit più significativo pari a zero

Char	Code	Char	Code	Char	Code	Char	Code	Char	Code	Char	Code	Char	Code	Char	Code
(NUL)	0	(DLE)	16	(space)	32	0	48	@	64	P	80	`	96	p	112
(SOH)	1	(DC1)	17	!	33	1	49	A	65	Q	81	a	97	q	113
(STX)	2	(DC2)	18	"	34	2	50	B	66	R	82	b	98	r	114
(ETX)	3	(DC3)	19	#	35	3	51	C	67	S	83	c	99	s	115
(EOT)	4	(DC4)	20	$	36	4	52	D	68	T	84	d	100	t	116
(ENQ)	5	(NAK)	21	%	37	5	53	E	69	U	85	e	101	u	117
(ACK)	6	(SYN)	22	&	38	6	54	F	70	V	86	f	102	v	118
(BEL)	7	(ETB)	23	'	39	7	55	G	71	W	87	g	103	w	119
(BS)	8	(CAN)	24	(	40	8	56	H	72	X	88	h	104	x	120
(HT)	9	(EM)	25	)	41	9	57	I	73	Y	89	i	105	y	121
(LF)	10	(SUB)	26	*	42	:	58	J	74	Z	90	j	106	z	122
(VT)	11	(ESC)	27	+	43	;	59	K	75	[	91	k	107	{	123
(FF)	12	(FS)	28	,	44	<	60	L	76	\	92	l	108	|	124
(CR)	13	(GS)	29	-	45	=	61	M	77	]	93	m	109	}	125
(SO)	14	(RS)	30	.	46	>	62	N	78	^	94	n	110	~	126
(SI)	15	(US)	31	/	47	?	63	O	79	_	95	o	111	(DEL)	127

Estensioni ANSI del codice ASCII ad 8 bit

- In un calcolatore, un carattere rappresentato mediante il codice ASCII è rappresentato su 8 bit (1 byte)
  - I code-point compresi tra 0 e 127 sono associati ai corripondenti code-point definiti dal codice ASCII
  - I restanti 128 code-point (128-255) sono utilizzati per rappresentare caratteri non compresi nel codice ASCII e specifici di una determinata lingua

- Nel tempo sono state definite diverse estensioni del codice ASCII su 8 bit
- Microsoft introdusse nel s.o. Windows diverse estensioni basate su una bozza di standard dell'ente americano ANSI
- Ciascuna di queste estensioni (dette ***code-page***) è identificata da un identificativo
  - Western European (CP1252)
  - Central European (CP1250)
  - Cyrillic (CP1251)
  - Greek (CP1253)
  - Hebrew (CP1255)
  - Arabic (CP1256)
  - …

Estensioni ISO del codice ASCII ad 8 bit

- Successivamente, l'ente di standardizzazione internazionale ISO definì nuovi set di caratteri ad 8 bit, denominati ISO-8859-*n*, con *n* = 1, 2, 3, …, 16
  - Il set di caratteri ISO-8859-1 (anche detto *Latin-1*) è identico a CP1252 ad eccezione dei code-point compresi tra 128 e 159, che sono usati per caratteri stampabili in CP1252 e per caratteri di controllo in ISO-8859-1
  - Il code-point 200 corrisponde a Č in ISO-8859-2, ed alla lettera cirillica Ш in ISO-8859-5

Lo standard Unicode

- La necessità di produrre documenti digitali in grado di contenere caratteri di diversi alfabeti ha condotto fin dagli anni ’90 alla definizione di ***Unicode***
- Unicode è definito dall’*Unicode Consortium* e sostanzialmente coincide con lo *Universal Character Set* (UCS) definito da ISO/IEC 10646
- Unicode codifica i caratteri usati in quasi tutte le lingue vive ed in alcune lingue morte, nonché simboli matematici e chimici, l'alfabeto Braille, ideogrammi ecc.
- In Unicode un carattere è rappresentato da un code-point numerico
- I code-point Unicode attualmente definiti sono rappresentati in esadecimale dai numeri compresi tra 0x000000 e 0x10FFFD
- La rappresentazione dei caratteri in Unicode è tale che:
  - i primi 128 code-point coincidono con quelli definiti dal codice ASCII
  - i primi 256 code-point coincidono con quelli definiti da ISO-8859-1
  - i codepoint compresi tra 0x000000 e 0x00FFFF costituiscono il cosiddetto *Basic Multilingual Plane* (BMP) che comprende i caratteri usati da quasi tutte le lingue moderne e un grande numero di caratteri speciali
- Unicode definisce soltanto il significato dei code-point, ma non definisce come i code-point debbano essere memorizzati in un computer o in un file
- Sono possibili diverse tecniche di rappresentazione dei code-point Unicode
- La tecnica UTF-32 (detta anche UCS-4) rappresenta ciascun code-point su 4 byte
  - Questa tecnica comporta un’occupazione di memoria quadrupla rispetto ad ASCII

La tecnica UTF-8

- La tecnica UTF-8 (*Unicode Transformation Format*) rappresenta i code-point Unicode mediante un codice a lunghezza variabile
- Essa rappresenta:
  - i 128 caratteri di codice Unicode compreso tra 0x000000 e 0x00007F (caratteri ASCII) mediante un solo byte che in binario ha la configurazione:
    0XXXXXXX
  - i 1920 caratteri di codice Unicode compreso tra 0x000080 e 0x0007FF, mediante una sequenza di due byte consecutivi che in binario hanno la configurazione:
    110XXXXX 10XXXXXX
  - i caratteri di codice Unicode compreso tra 0x000800 e 0x00FFFF mediante una sequenza di tre byte consecutivi che in binario hanno la configurazione:
    1110XXXX 10XXXXXX 10XXXXXX
  - i restanti caratteri, di codice Unicode compreso tra 0x010000 e 0x10FFFD, mediante quattro byte consecutivi: 11110XXX 10XXXXXX 10XXXXXX 10XXXXXX
- Si osservi che se un carattere è codificato da una sequenza di *n* byte con *n*>1:
  - il primo byte della sequenza con i suoi primi n+1 bit indica la lunghezza della sequenza
    - 110 → due byte, 1110 → tre byte, 11110 → quattro byte
  - i byte successivi al primo hanno sempre 10 come bit più significativi
- Il codice UTF-8 gode della proprietà del prefisso: nessuna sequenza di byte corrispondente a uno specifico carattere è contenuta in una sequenza più lunga che codifichi un carattere diverso
  - se uno o più byte andassero persi, sarebbe possibile risincronizzare la decodifica all'inizio del carattere successivo, limitando l'effetto dell'errore

La tecnica UTF-16

- La tecnica UTF-16 rappresenta i code-point Unicode mediante un codice a lunghezza variabile
- Essa rappresenta:
  - i 65536 caratteri del BMP, aventi codice Unicode compreso tra 0x000000 e 0x00FFFF, mediante una sequenza di due byte (16 bit) che ne rappresenta il valore
  - i restanti caratteri, di codice Unicode compreso tra 0x010000 e 0x10FFFD mediante una coppia “surrogata” di sequenze di due byte (per un totale di 32 bit) costruite a partire dal codice Unicode del carattere in modo da assumere entrambe un valore compreso tra 0xD800 e 0xDFFF
- Ad esempio:
  - il carattere ‘z’ di codice Unicode (ed ASCII) 122=0x7A
    si rappresenta in UTF-16 come 0x007A (16 bit)
  - l’ideogramma cinese 水 (acqua) di codice Unicode 27700=0x6C34
    si rappresenta in UTF-16 come 0x6C34 (16 bit)
  - Il carattere 𝄞 (chiave di Sol) di codice Unicode 119070=0x1D11E
    si rappresenta in UTF-16 come 0xD834 0xDD1E