💩.la?!? the fascinating history and current state of IDNs!!

Hello

I'm @KTamas

I'm here to talk about IDNs
(Internationalized Domain Names)

DNS

- Translates `www.facebook.com` to `31.13.84.36`
- Is limited to just 37 characters
- Letters, numbers, dashes
- Yup, that's it
- More importantly, 63 characters for each part of the domain (separated by the .)
- 255 characters total, with the dots

37 characters should be enough for ev-

Why should we care about IDNs?

- Not everyone can read the Latin alphabet
- People have the right to use their own languages everywhere
- There are a lot of languages using the Latin alphabet that use more than 26 characters
- …

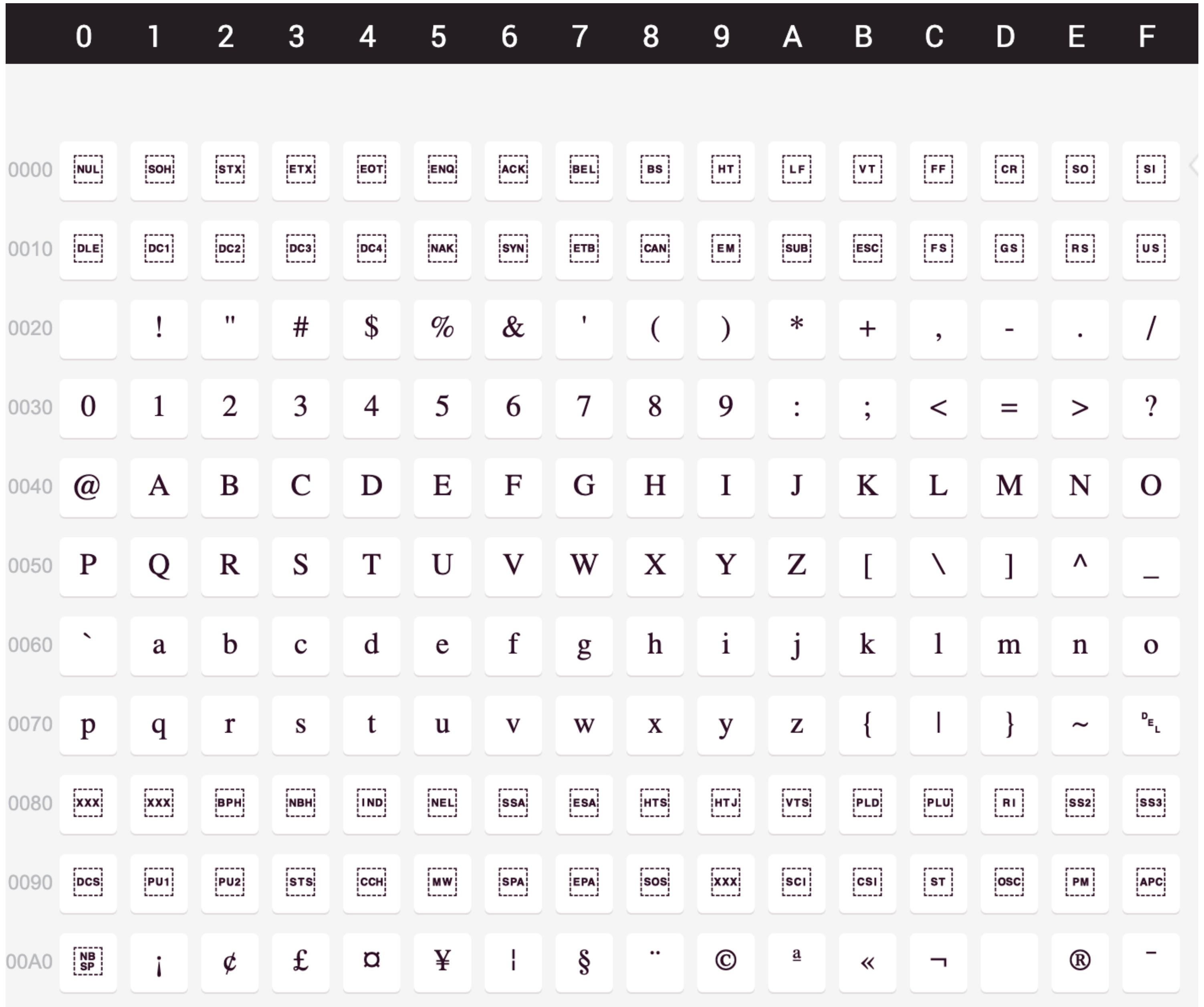
History, part 1

- Mid 80s: DNS
- 1989: World Wide Web
- 1991: Mosaic
- you know the rest

Detour: Unicode

- A big table (array) of characters, ideograms, emojis etc.
- Over 1 million code points
- Usually written as U+XXXX (where XXXX is in hex, and sometimes it's more than 4 characters)

Unicode-table.com



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0000	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
0010	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
0020		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
0030	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
0040	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
0050	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
0060	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
0070	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL
0080	XXX	XXX	BPH	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3
0090	DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	XXX	SCI	CSI	ST	OSC	PM	APC
00A0	NB SP	¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬		®	¯

History, part 2

- 1987: Work on Unicode starts
- 1991: Unicode 1.0
- 1993: UTF-8 — variable-width encoding for the masses
- 1995: `<meta charset="">` in Netscape
- 1996: Unicode support in Netscape
- 2001: Windows XP, the first fully-Unicode consumer Windows
- 2010: Emojis in Unicode 6.0

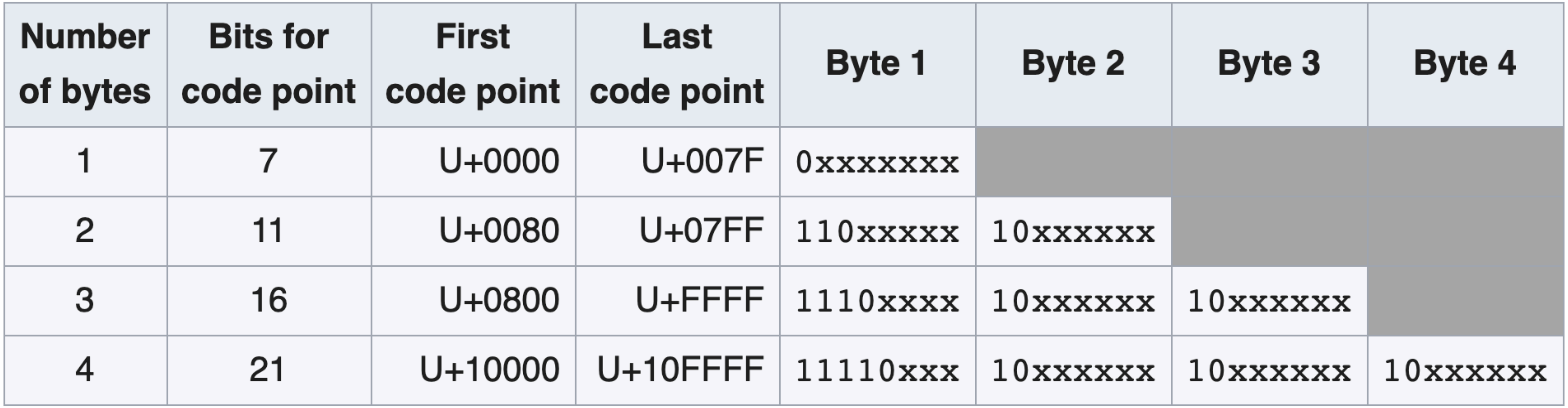
History, part 3

- Today we're at Unicode 12.1
- UTF-8 is everywhere
- You don't have to manually set your encoding in your browser

Back to the mid-90s

- Most of the web is in English
- ISO-8859 aka Latin 1 or find your own encoding
- Which is what people did and it was a mess
- UTF-8 mostly fixed this
- But we need something else for domains: maybe UTF-5?

The case for variable-width encodings: UTF-8



Number of bytes	Bits for code point	First code point	Last code point	Byte 1	Byte 2	Byte 3	Byte 4
1	7	U+0000	U+007F	`0xxxxxxx`			
2	11	U+0080	U+07FF	`110xxxxx`	`10xxxxxx`		
3	16	U+0800	U+FFFF	`1110xxxx`	`10xxxxxx`	`10xxxxxx`	
4	21	U+10000	U+10FFFF	`11110xxx`	`10xxxxxx`	`10xxxxxx`	`10xxxxxx`

1996: The journey begins with UTF-5

- by Martin Dürst
- Everything needs to fit into existing constraints
- 37 characters to pick from, 63 characters for each part, 255 characters in total
- That's not a lot, so let's get creative

is.s.u-tokyo.ac.jp

information.science.university-of-tokyo.academia.japan

情報.り.東大.学.日本

jouhou.ri.toudai.gaku.nihon

情報.り.東大.学.日本

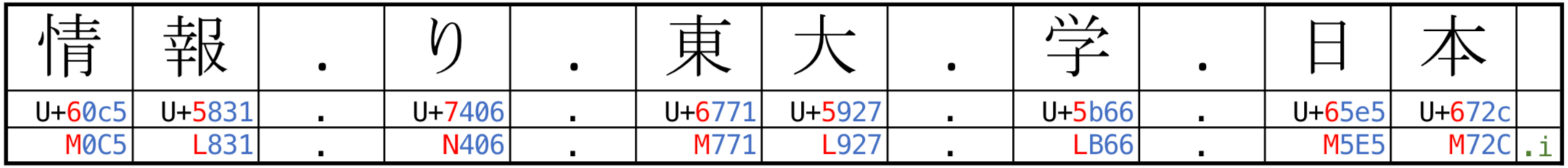
U+60c5U+5831.U+7406.U+6771U
+5927.U+5b66.U+65e5U+672c

M0C5L831.N406.M771L927.LB66.
M5E5M72C.i

UTF-5

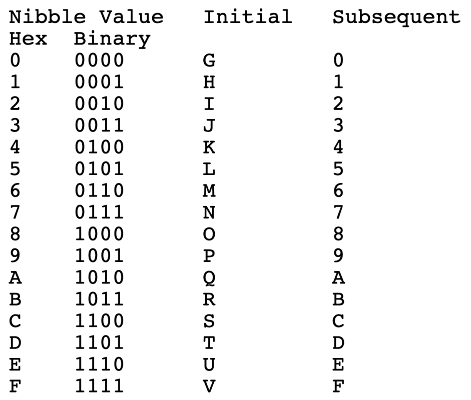
- The same(ish), but with only 32 characters (2^5)
- Example domain: is.s.u-tokyo.ac.jp
- Which means: information.science.university-of-tokyo.academia.japan
- In Japanese: 情報.り.東大.学.日本
- Transliterated: jouhou.ri.toudai.gaku.nihon
- Unicode: U+60c5U+5831.U+7406.U+6771U+5927.U+5b66.U+65e5U+672c
- UTF-5: M0C5L831.N406.M771L927.LB66.M5E5M72C.i

UTF-5, Illustrated



情	報	.	り	.	東	大	.	学	.	日	本	
U+60c5	U+5831	.	U+7406	.	U+6771	U+5927	.	U+5b66	.	U+65e5	U+672c	
M0C5	L831	.	N406	.	M771	L927	.	LB66	.	M5E5	M72C	.i

- U+60c5 becomes M0c5 (6 becomes M), U+5831 becomes L831 (5 becomes L)...
- Variable-length: U+0234 becomes I34 etc.



Nibble Value		Initial	Subsequent
Hex	Binary		
0	0000	G	0
1	0001	H	1
2	0010	I	2
3	0011	J	3
4	0100	K	4
5	0101	L	5
6	0110	M	6
7	0111	N	7
8	1000	O	8
9	1001	P	9
A	1010	Q	A
B	1011	R	B
C	1100	S	C
D	1101	T	D
E	1110	U	E
F	1111	V	F

UTF-5 is a good start, but…

- It's limited
- ~15 ideographs
- ~21 Hebrew or Arabic characters
- Can't mix with latin characters (those would have to be encoded as well)
- It's a good start though, so let's get to work

1996-2003

- 1998: Working Group was formed
- 1999: Several test implementations
- 1999: Taiwan: `.gongsi` aka `.公司` aka `.com`; 200k sold
- 1999: India: Tamil versions of `.com`/`.net`/`.org`/`.edu`
- 2001: ICANN IDN Committee
- 2003: RFC 3454, RFC 3490, RFC 3491 and RFC 3492!

RFC-3492 aka Punycode

- UTF-5 on steroids
- `xn--whatever-adsf7u347q34.com`

RFC-3492 aka Punycode

- Why `xn--`?
- ACE (ASCII Compatible Encoding)
- "Hey, punycode domain incoming" to the Browser

RFC-3492 aka Punycode

- Punycode is not pretty, but very efficient
- Let's assume a domain is likely in one language
- Those language's code points are usually next to each other
- Let's separate the characters into two groups: the OG 37 and the rest
- Take the lowest code point from the rest, encode it
- For the rest, just encode the distance between them
- Encode the location of the character as well within the whole string

Examples!

- bücher.example -> xn--bcher-kva.example
- 💩.la -> xn--ls8h.la
- ☃.net -> xn--n3h.net
- hello🇩🇪.com -> xn--hello-my73dha.com

Examples!

- `apple.co` -> `xn--le-6kc8da.xn--n1af`

Examples!

- `apple.co` -> `xn--le-6kc8da.xn--n1af`
- 'a', 'p', 'c', 'o' are Cyrillic homographs
- We have a problem

IDN homograph attacks

- Huge problem
- Used for phishing etc.
- Because of this, browsers don't always display the IDN, but punycode instead
- Complex algorithms decide when to display what
- On mobile it's even worse since it's common to hide the URL in the UI

Where are we at right now?

- Several dozen TLDs
- Emoji domains available for 14 TLDs (like `.ws`)
- 7.5 million IDN domains, 2% of all domains
- IDN emails are a thing but support is still scarce

Thank you

- Twitter: @KTamas
- This presentation: http://i❤️idns.ktamas.com/ (http://xn--iidns-102c.ktamas.com/)
- More resources, links, references on the next two slides

Resources, links

- Original UTF-5 Draft
- IDNs - Wikipedia
- A presentation from 2004 with lots of details
- iDNS project/APNG commission (didn't have time for this)
- The best explanation for Punycode
- RFC 3492
- Variable-width encoding
- IDN history

Resources, links

- List of TLDs, including Internationalized ones
- Emoji domains - Wikipedia
- Punycoder - convert from/to Punycode
- Register emoji domains
- IDN World Report