情報処理システム論 (13)

文字コードとは

- 文字の符号化方法
- ISO(国際標準化機構)にる国際規格
  - 「情報交換用の」(コンピュータ間の通信など)
  - たとえばASCII
    - ANSI(米国規格協会)による規格
    - ISO646 BCT(Basic Code Table) 国際標準

ISO646 BCT(Basic Code Table)

- **ASCII と異なってもよい 12 文字**

2/3	2/4	4/0	5/11	5/12	5/13	5/14	6/0	7/11	7/12	7/13	7/14
#	$	@	[	\	]	^	`	{	|	}	~
#	$	@	[	¥	]	^	`	{	|	}	‾
£	$	@	[	\	]	^	`	{	|	}	~

上が ASCII
中央が JIS X0201ローマ字
下が BSI 4730

文書とストリーム

- 文書を構成する文字列はストリームとして扱われる

ABCDEFG
こんにちは
12345

ABCDEFG[改行]こんにちは[改行]12345

文字コードの拡張方法

- 複数の文字セットを混在させるため

  5 ~　(3/5) (7/14)

  京　(3/5 7/14)

を区別したい

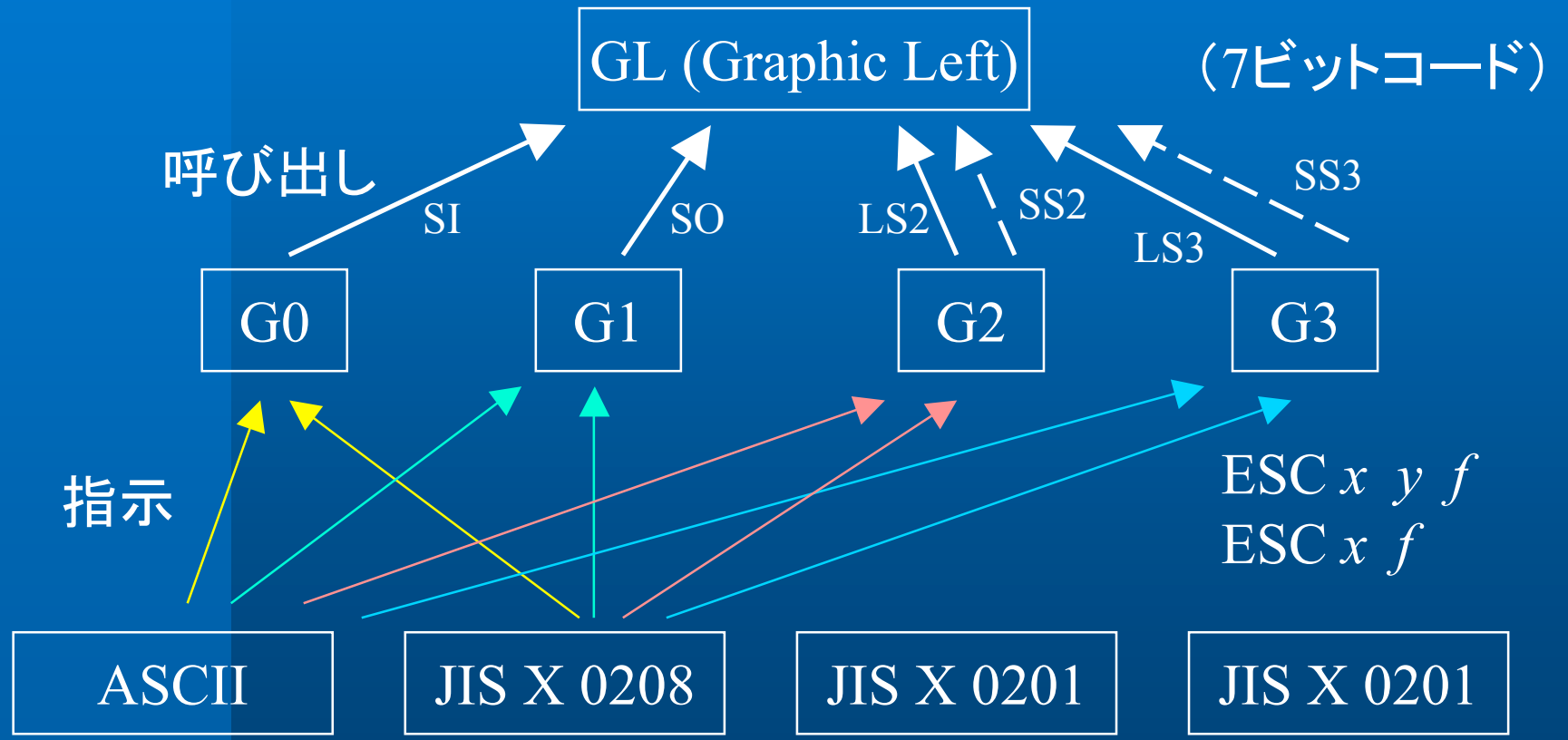
- 特殊なコード列をはさむことで区別する
  - ISO-2022 (情報交換符合の拡張法)
  - JIS X 0202

ISO-2022 による拡張

- 7単位符合の拡張法
- 8単位符合の拡張法

7単位符合の拡張法

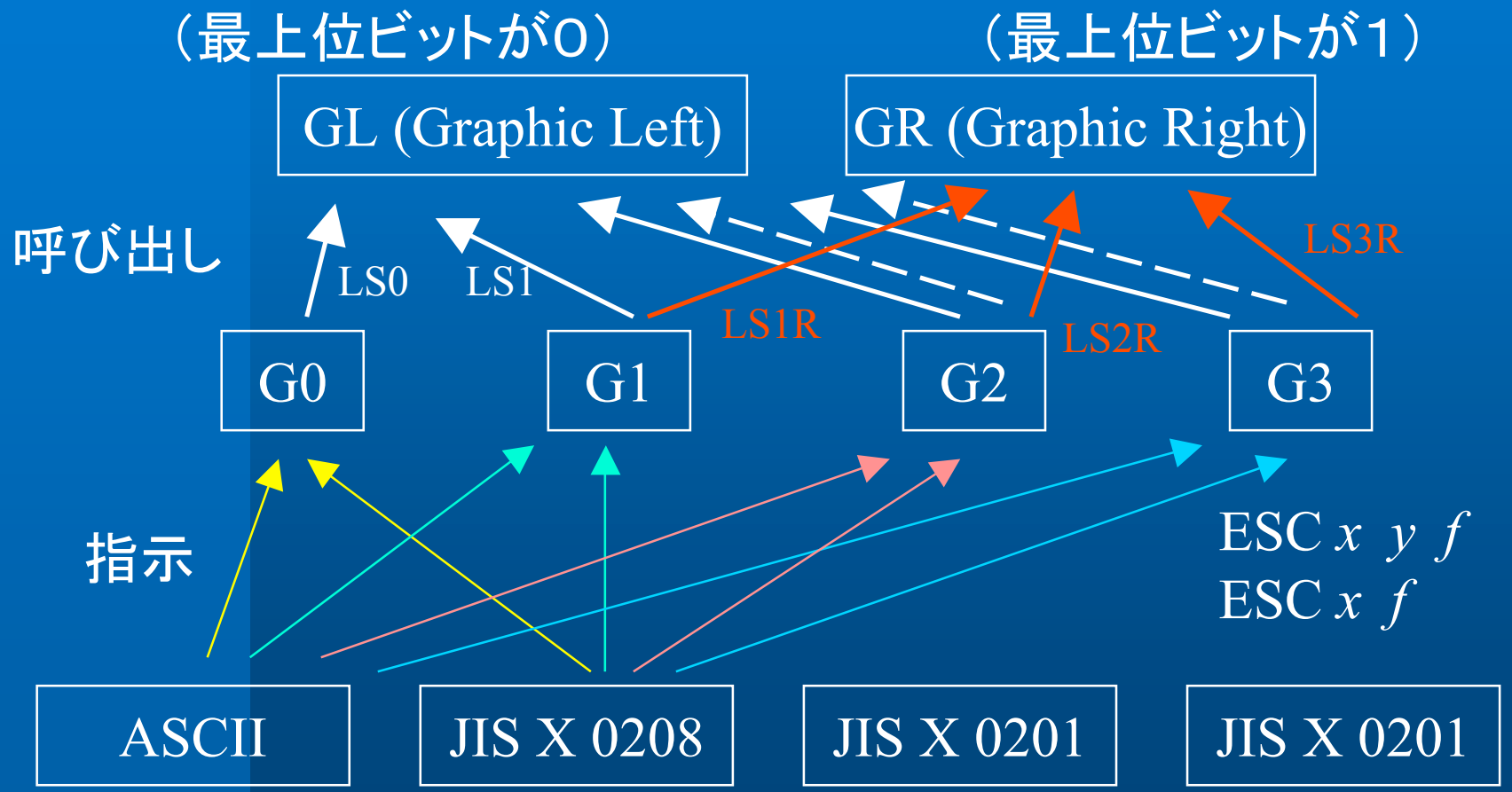
Shift In
Shift Out
Locking Shift
Shingle Shift



ISO-2022-JP の場合

- G0 を GL に呼び出したままにしておく
- ASCII を G0 に指示
  - ESC 2/8 2/4 (ESC (B)
- 漢字(JIS X 0208)を G0 に指示
  - ESC 2/8 4/0 (ESC $ B)
- ローマ字(JIS X 0201) を G0 に指示
  - ESC 2/8 4/10 (ESC (J)

8単位符合の拡張法


（最上位ビットが0）
（最上位ビットが1）
GL (Graphic Left)
GR (Graphic Right)
呼び出し
LS0
LS1
LS1R
LS2R
LS3R
G0
G1
G2
G3
指示
ESC x y f
ESC x f
ASCII
JIS X 0208
JIS X 0201
JIS X 0201

EUC-Japan の場合

- JIS X 0201 ローマ字を G0 に固定的に指示
- JIS X 0208 漢字を G1 に固定的に指示
- JIS X 0201 カナを G2 に固定的に指示
- JIS X 0212 補助漢字を G3 に固定的に指示
- 基本
  - G0 を GL に固定的に呼び出し
  - G1 を GR に固定的に呼び出し
  - G2, G3 は Single Shift で呼び出し

文字コードの統一

- 欧米
  - 1バイト文字圏
  - 8ビットコードの混在
    - 8ビット目の立った部分
- アジア
  - 2バイト文字圏
  - 16ビットコードの混在

ISO 10646 と UNICODE

- 地域化作業の負担の軽減
  - 一つ作ればどこででも使える
- 多言語化も目指す?
- すべてを 2 または 4 バイトコードに割付
  - 英数字もいっしょ
- JIS X 0221-1995

  Information technology - Universal Multiple-Octet Coded Character Set (UCS) - Part 1 : Architecture and Basic Multilingual Plane

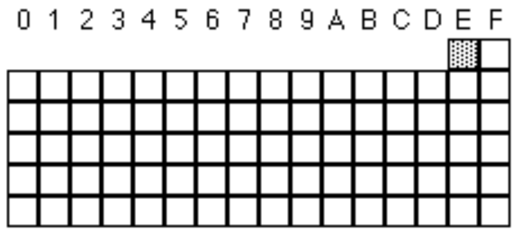
UCS

- UCS-2 (UNICODE)
  - 1文字2バイト
- UCS-4
  - 1文字4バイト
  - 00群00面 ー Basic Multilingual Plane (BMP) として UNICODE を採用

UCS-2 の4つの領域

- A領域 (0000 - 4DFF)
  - ラテン、ギリシャ、キリル、アラビア、ヘブライ等のアルファベットとタイ語、ハングル（非表意文字）
- I領域 (4E00 - 9FFF)
  - 漢字、Hanji（中国語）、Hanja（韓国語）
  - 統一CJK表意文字 (unified CJK ideo graphs)
- O領域（将来のために予約）
- R領域 (E000 - FFFD)
  - 私用文字、互換文字など

U+4E00 - 4EFF



U+4E00 - U+4EFF

To change to another block, click on one of the boxes to the left.

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	一	丐	丠	丰	乀	乐	习	买	亀	亐	亠	亰	什	仐	仠	仰
1	丁	丑	両	丱	乁	乑	乡	乱	亁	云	亡	亱	仁	仑	仡	仱
2	丂	丒	丢	串	乂	乒	乢	乲	亂	互	亢	亲	仂	仒	仢	仲
3	七	专	丣	丳	乃	乓	乣	乳	亃	亓	亣	亳	仃	仓	代	仳
4	丄	且	两	临	乄	乔	乤	乴	亄	五	交	亴	仄	仔	令	仴
5	丅	丕	严	丵	久	乕	乥	乵	亅	井	亥	亵	仅	仕	以	仵
6	丆	世	並	丶	乆	乖	书	乶	了	亖	亦	亶	仆	他	仦	件
7	万	丗	丧	丷	乇	乗	乧	乷	亇	亗	产	亷	仇	仗	仧	价
8	丈	丘	丨	丸	么	乘	乨	乸	予	亘	亨	亸	仈	付	仨	仸
9	三	丙	丩	丹	义	乙	乩	乹	争	亙	亩	亹	仉	仙	仩	仹
A	上	业	个	为	乊	乚	乪	乺	亊	亚	亪	人	今	仚	仪	仺
B	下	丛	丫	主	之	乛	乫	乻	事	些	享	亻	介	仛	仫	任
C	丌	东	丬	丼	乌	乜	乬	乼	二	亜	京	亼	仌	仜	们	仼
D	不	丝	中	丽	乍	九	乭	乽	亍	亝	亭	亽	从	仝	仭	份
E	与	丞	丮	举	乎	乞	乮	乾	于	亞	亮	亾	仍	仞	仮	仾
F	丏	丟	丯	丿	乏	也	乯	乿	亏	亟	亯	亿	仏	仟	仯	仿

UTC-2 の文字の例

ヘブライ

ハングル

アラビア

その他の記号

タイ

インターネットとUNICODE

- インターネットでは8ビットコードは通らない
  - 歴史的経緯とbackward compatibility
- RFC1642
  - UTF-7 A Mail-Safe Transformation Format of Unicode
- RFC2044
  - UTF-8, a transformation format of Unicode and ISO 10646
- UTF ― UCS Transformation Format

charset=UNICODE-1-1

- **RFC1641 : Using Unicode with MIME**
- **メールの本文に UCS-2 をそのまま書くとき**

- **「日本語」とかいた場合の例**

Content-Type: text/plain; charset=UNICODE-1-1
Content-Transfer-Encoding: base64

ZeVnLIqe

UTF-7

- ASCII はそのまま
- UNICODE の2バイトを base64 エンコードして、+ と - ではさむ
- 場合によって、- は省略可能
- + は +- と表記

- A≠α (0041, 2262, 0391, 002E) → A+ImIDkQ.
- 日本語 (65E5, 672C, 8A9E)→ +ZeVnLIqe-

UFT-8

- **ASCII の 1バイト表現を保ちながら UCS-4 を利用する方法**
- **A≠α (0041, 2262, 0391, 002E)**

  → 41 E2 89 A2 CE 91 2E

UCS-4 range (hex.)	UTF-8 octet sequence (binary)
0000 0000-0000 007F	0xxxxxxx
0000 0080-0000 07FF	110xxxxx 10xxxxxx
0000 0800-0000 FFFF	1110xxxx 10xxxxxx 10xxxxxx
0001 0000-001F FFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx
0020 0000-03FF FFFF	111110xx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx
0400 0000-7FFF FFFF	1111110x 10xxxxxx … 10xxxxxx

参考文献

- いま日本語が危ない
  ～文字コードの誤った国際化～
  太田昌孝, 丸山学芸図書
  ISBN 4-89542-146-5, \2000
- http://turbine.kuee.kyoto-u.ac.jp/FAQ/kanji-code.html
- http://www.unicode.org