中文世界

darkx

編碼標準 Encoding Standard

- 電腦是美國人發明的
  - ASCII (American Standard Code for Information Interchange)

- 地方的電腦也要顯示中文
  - Big5
  - 台灣財團法人資訊工業策進會 在 1983 年為 五大中文套裝軟體 設計的編碼系統
  - 繁體中文中最常用的電腦中文字符集標準
  - 萬年遺毒

編碼標準

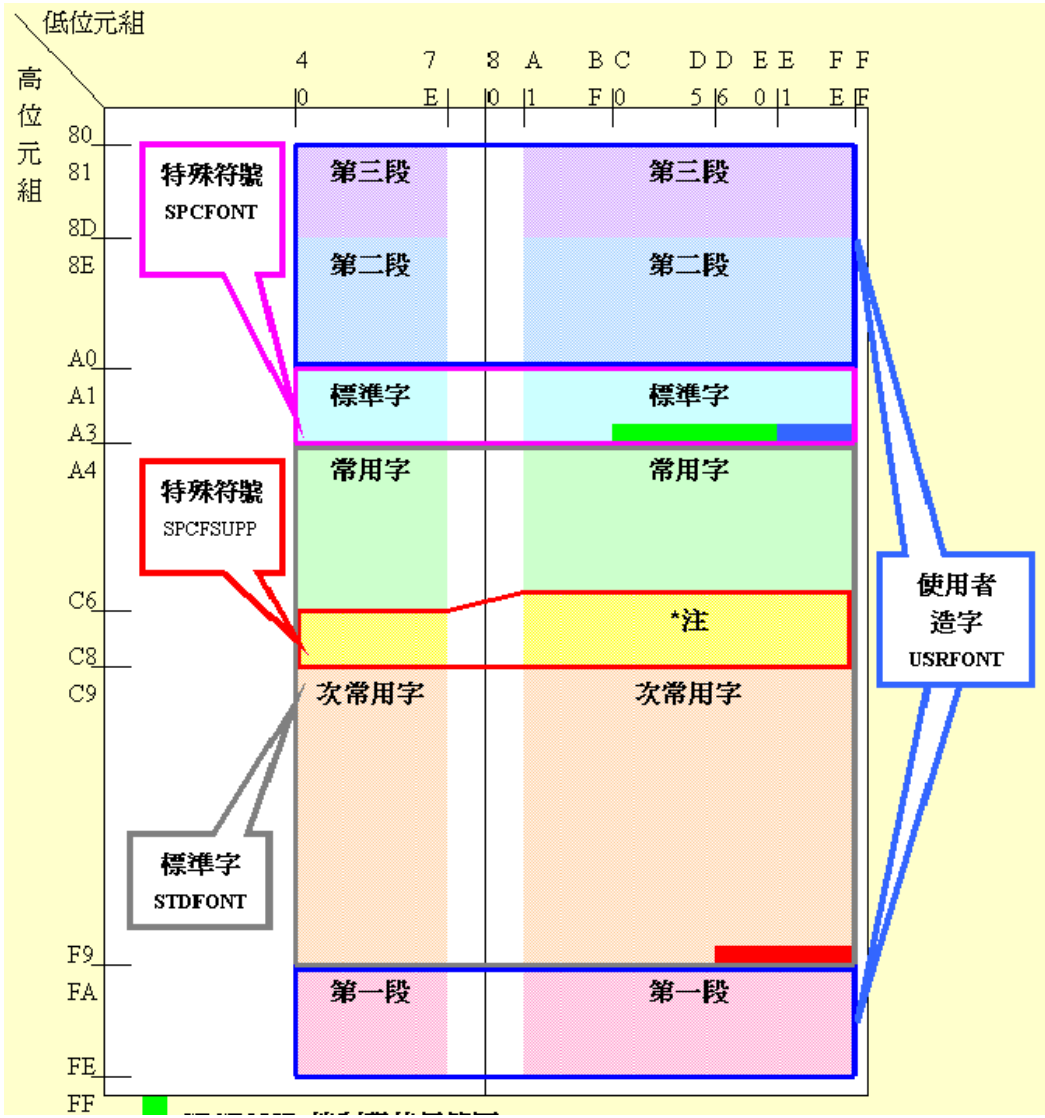
- ASCII
  - 8 bits (256 個組合)
  - 實際上指 0x00 ~ 0x7F 共 128 種字元
    - 0x00 ~ 0x1F：control characters
    - 0x20 ~ 0x7E：printable characters
    - 0x7F: delete
  - 0x80~0xFF：Extension
  - man ascii

- Big5
  - 使用 2 bytes 來存放中文字 (理論上有 65536 種可能)
  - 為了與 ASCII 相容, 只能使用 19782 個
    - [0x81 ~ 0xFE][0x40~0x7E, 0xA1 ~ 0xFE]
      = 126 * (63 + 94) = 126 * 157 = 19782

Ref: http://www.cns11643.gov.tw/AIDB/encodings.do

編碼標準 – Big5

- 標準字 (13503 個)
  - 常用字
    - 你我他的媽
  - 次常用字
    - 枋晃束鐰飉
- 特殊符號 (441 個)
  - 符號、控制碼
    - ：！。∩♂├
  - 罕用符號
- 使用者造字區 (5809 個)
  - 分為三段



Big5 的問題

- 使用者造字區
  - 每個人都可以自己造字, 自己造的字放到別人電腦上看不到
- 缺字
  - 塋、煊、栢、喆
- 各家實作延伸版本繁雜
  - 倚天中文 Big5 延伸
  - Code Page 950 " Big5 事實標準"
  - Big5+
  - Big-5E
  - Big5-2003
- 許功蓋問題
  - 0x5C (\) 會有特殊意義
  - 許 (0xB35C) 功 (0xA55C) 蓋 (0xBB5C)

編碼標準 - Unicode

全世界共有上百種文字, 因此有很多種不同的編碼系統

日本有 **JIS**, 中國有 **GB 2312**, ... etc

同樣的編碼在不同的編碼系統下顯示會不同

Unicode 組織就誕生了

Unicode V.S. ISO 10646

- 1991 年左右, 同時有兩個組織著手規範世界通用編碼
  - Unicode
  - ISO 10646

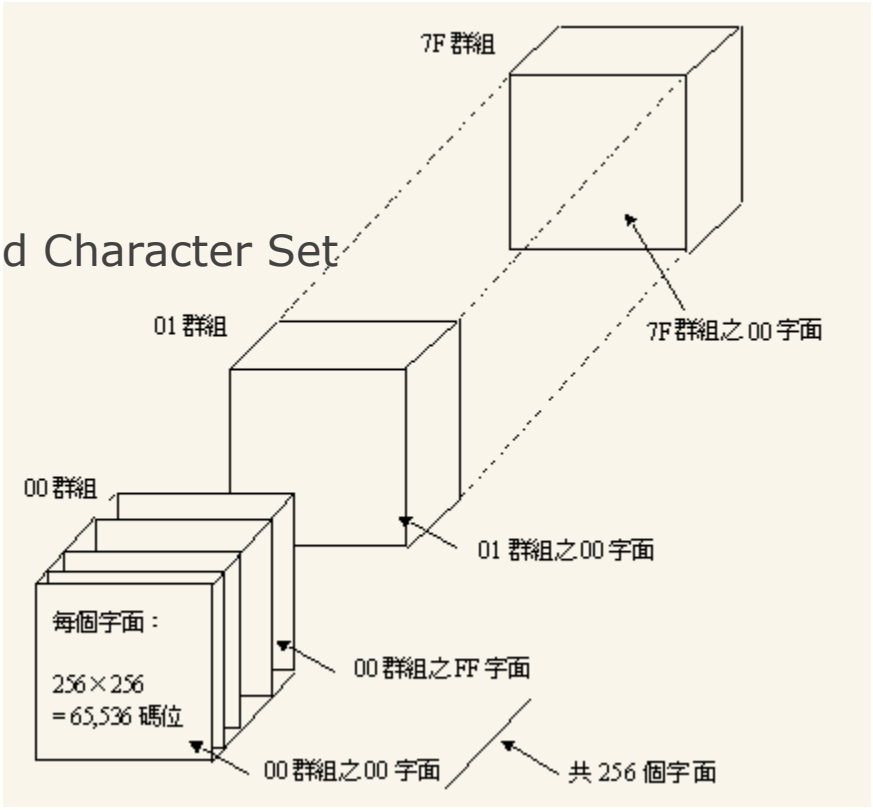
「這個世界不需要兩套不同的單一字符集」
  - 因此他們決定共用同樣的字碼

- 現在這兩個組織各自存在, 各自互相砥礪

ref:http://zh.wikipedia.org/wiki/ISO_10646

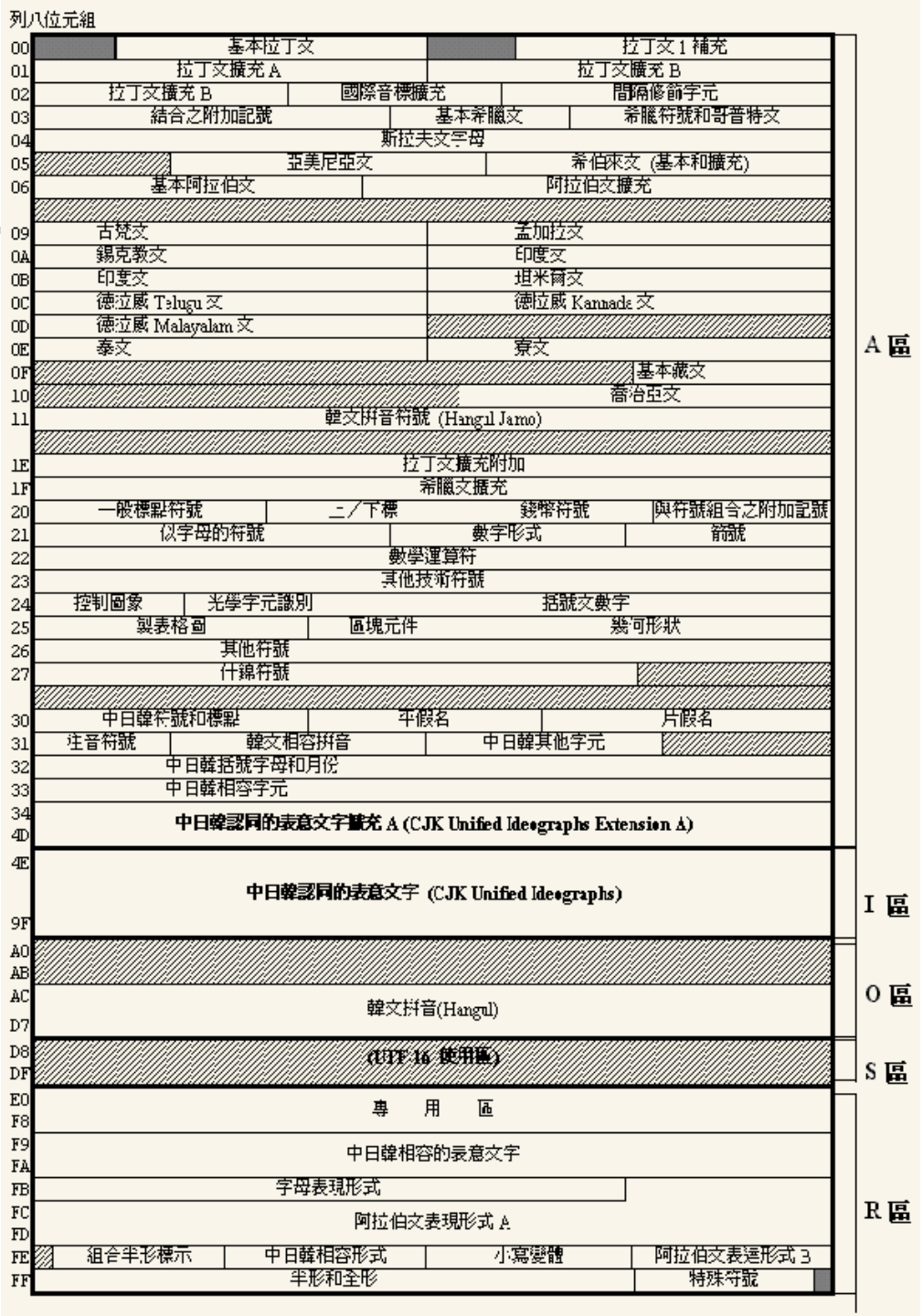
編碼標準 – ISO10646 and Unicode (1)

- Goal
  - 集結全球通用字符集的聯集
  - UCS-4
    - Universal multiple-octet coded Character Set
  - 4 bytes encoding (2^31)
    - 128 Groups
    - 256 Planes each group
    - 256 Rows each plane
    - 256 Cells each row
  - BMP (UCS-2)
    - Basic Multilingual Plane
    - 00 group, 00 plane
    - 65536 encoding space
  - Why in BMP
    - **若所有字集都在 BMP 中, 就可以只使用 2 bytes, 否則就要用 4 bytes, 不能混用**



編碼標準 – ISO10646 and Unicode (2)

- BMP



Unicode 的問題

- Big Endian & Little Endian
  - U+4E59？(乙)
  - U+594E？(奎)

- 編碼空間浪費
  - ASCII 字元通通都用 2byte 表示:0x00 0x41「A」第一位永遠是0

Unicode Transformation Format

- UTF: UCS/Unicode Transformation Format
  - UTF-16 (2、4 bytes)
    - 將一個 32-bit ISO10646 字元轉成多個 16-bit Unicode
    - Windows
  - UTF-8 (1~4 bytes)
    - 將一個 32-bit ISO10646 字元轉成多個 8-bit Unicode
    - 將一個 16-bit Unicode 字元轉成多個 8-bit Unicode
    - 128 個 US-ASCII 字元只需 1 bytes 編碼
    - 帶有附加符號的拉丁文、希臘文、西里爾字母、亞美尼亞語、希伯來文、阿拉伯文、敘利亞文及它拿字母則需要 2 bytes 編碼
    - 其他基本多文種平面(BMP)中的字元(這包含了大部分常用字)使用 3 bytes 編碼
    - 其他極少使用的 Unicode 輔助平面的字元使用 4 bytes 編碼
    - Unix-like systems

中文環境 (1)

- 要做到哪些事情
  - 中文訊息
  - 中文顯示
  - 中文輸入
  - 中文列印
  - 中文處理

簡單

困難

中文環境 (2)

- 中文化方式
  - 直接修改程式
    - 成千上萬的程式只有真強者才能改完了
      18 chars
  - 國際化(InternationalizatioN, i18n)
    - Multi-language architecture
      - 程式設計人員按照該架構的機制與準則寫程式, 便可支援各式各樣的語言
    - Locale (LOCALization Environment database)
      - 程式根據使用者選擇的 locale 聯繫到不同資料庫, 進而提供該語言的支援
  - 在地化(LocalizatioN, L10n)
    - 在 i18n 的大架構下, 加入「在地化」的特性
  - 通常 i18n 只需做一次, 而 L10n 要針對每個語言個別做

i18n & L10n

- i18n + L10n
  - 語言/翻譯
  - 文化、書寫習慣
    - 名字和稱謂的位置
    - 電話號碼, 位址和國際郵遞區號的格式
    - 貨幣單位
    - 度量衡
    - 日期時間
    - 時區
    - 數字格式
- L10n only
  - 內容在地化
  - 道德在地化
  - 文化價值
  - 社會環境

中文環境 (3)

- locale in FreeBSD
  - 地區性語言的資訊
    - LC_ALL
    - LC_CTYPE (字元的處理方式)
    - LC_MESSAGES (程式訊息輸出所用的語言)
    - LC_TIME (時間格式)
    - LC_NUMERIC (數字格式)
    - LC_MONETARY (貨幣格式)
    - LC_COLLATE (字母順序與特殊字元比較)
    - LANG (語言顯示)
  - 效力優先性:LC_ALL > LC_* > LANG

中文環境 (4)

- 設定 locale
  - csh/tcsh shell
    - `setenv LC_CTYPE  en_US.UTF-8`
  - Bourne Shell
    - `export LC_CTYPE=en_US.UTF-8`

  Note: 可以寫在 .tcshrc/.bashrc 中登入後自動載入

  - /usr/share/locale/
    - 各國的 locale 資訊
    - 命名規則:語言_地區名:字元編碼名稱
      - zh_TW.UTF-8
      - zh_CN.GBK

中文環境 (5)

- 中文 Terminal (Remote Login)
  - M$ Windows: putty, pietty …etc.
  - X Window: xterm, rxvt-unicode ,roxterm…etc.
  - 設定好中文支援, 登入後即可看到中文
    - setenv LC_CTYPE en_US.UTF-8 (csh/tcsh)
    - export LC_CTYPE=en_US.UTF-8 (sh/bash)
    - 顯示為英文但支援 multibyte characters

- 中文 Xwindow
  - 建立支援 L10n 中文環境
    - 安裝中文字型
    - 設定 Shell locale 環境
    - 安裝中文輸入法 (Ex. ibus)

中文世界 HOWTO (hw1-3)

Steps

- 安裝中文字型
- 安裝中文 Terminal Emulator
- 安裝中文輸入法 (Ex. ibus)
- 其他設定

中文字型

- 兩大中文字型種類
  - 點陣字型 (Bitmapped Font)
    - BDF (Bitmap Distribution Format) 點陣分散格式
    - HBF (Hanzi Bitmap Font) 漢字點陣字體
    - PCF (Portable Compiled Font)
  - 曲線描邊字型 (Outline Fonts)
    - True Type Font (TTF)

安裝中文字型 (2)

Font Path:
/usr/local/lib/X11/fonts/misc/,
/usr/local/lib/X11/fonts/TTF/,
/usr/local/lib/X11/fonts/Type1/,
/usr/local/lib/X11/fonts/75dpi/,
/usr/local/lib/X11/fonts/100dpi/,
/usr/local/lib/X11/fonts/local/

- Font Path
  - % xset q
  - % xset fp+ [directory]
  - % xset fp rehash

- 安裝字型
  - 1. 直接從 Windows 下偷過去
  - 2. 透過 ports 安裝字型檔案

  - 使用 fc-cache 建立字型資料庫
  - 修改各軟體設定使用別的字型

安裝中文字型 (3)

- 安裝 ttfm – TrueType Font Manager 字型管理工具
- ttfm
  - ttfinfo 讀取 ttf 字型格式資訊的程式
    - **% ttfinfo /usr/local/share/fonts/TrueType/fireflysung.ttf**

```
(21:38)wjguo@[oopc6:/home/wjguo] >ttfinfo /usr/local/share/fonts/TrueType/fireflysung.ttf
TTFINFO_FONT_FILE="/usr/local/share/fonts/TrueType/fireflysung.ttf"
TTFINFO_FACE_NUM="1"
TTFINFO_FACE_INDEX="0"
TTFINFO_FONT_NAME="AR PL New Sung"
TTFINFO_FONT_PSNAME="AR-PL-New-Sung"
TTFINFO_FOUNDRY_NAME="misc"
TTFINFO_WEIGHT_NAME="medium"
TTFINFO_WIDTH="normal"
TTFINFO_NUMCMAP="3"
TTFINFO_CMAP0="0,3"
TTFINFO_CMAPNAME0="Apple Unicode,(v.2.0)"
TTFINFO_CMAP1="1,0"
TTFINFO_CMAPNAME1="Apple,Roman"
TTFINFO_CMAP2="3,1"
TTFINFO_CMAPNAME2="Windows,Unicode"
TTFINFO_MAPNUM="3"
TTFINFO_FONTMAP1="-misc-AR PL New Sung-medium-r-normal--0-0-0-0-p-0-big5-0"
TTFINFO_FONTMAP2="-misc-AR PL New Sung-medium-r-normal--0-0-0-0-p-0-gb2312.1980-0"
TTFINFO_FONTMAP3="-misc-AR PL New Sung-medium-r-normal--0-0-0-0-p-0-jisx0208.1983-0"
```

安裝中文字型 (4)

- **ttfm.sh**

```
(21:38)wjguo@[oopc6:/home/wjguo] >ttfm.sh

True-Type Font Manager 0.9.3

Usage: /usr/local/bin/ttfm.sh [option]

      --add [module] <file>...    install ttf font
      --remove [module] <file>... remove ttf font from the system
      --list <module>... list all ttf fonts on the system
      --modules          list all ttf manager modules on the system
      --setdefault <module> <file>
                    set default ming font of module to file
      --setdefault_kai <module> <file>
                    set default kai font of module to file
      --initm <module>.. initialize modules
      --help             show this info
```

安裝中文字型 (5)

- 選一個來裝 ...

Chapter 6. 輸出字型

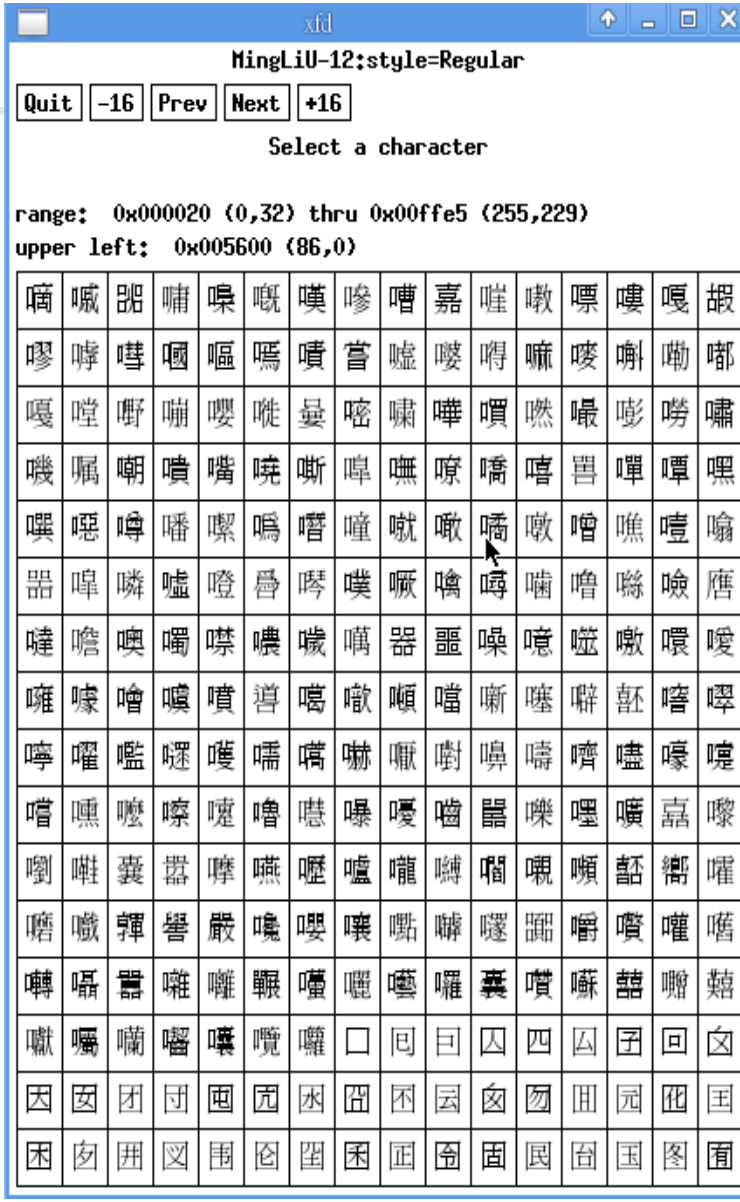
Table of Contents

安裝中文字型 (6)

- fireflyttf
  - 透過 ports 安裝的都會自己跑
    - % ttfm.sh --add xttfm /usr/local/share/fonts/TrueType/fireflysung.ttf
    - % fc-cache –f –v /usr/local/lib/X11/fonts/TrueType/
  - portmaster chinese/fireflyttf

*安裝中文字型 (7)

- 用 xfd 來看
  - % setenv LC_CTYPE zh_TW.Big5
  - % xfd –fa “MingLiU”
  - man xfd



安裝中文字型 (8)

- 增加 Font Path
  - Edit /etc/X11/xorg.conf
    - /usr/local/share/fonts/TrueType/fireflysung.ttf
    - /usr/local/lib/X11/fonts/TrueType/fireflysung.ttf
  - Restart xwindow

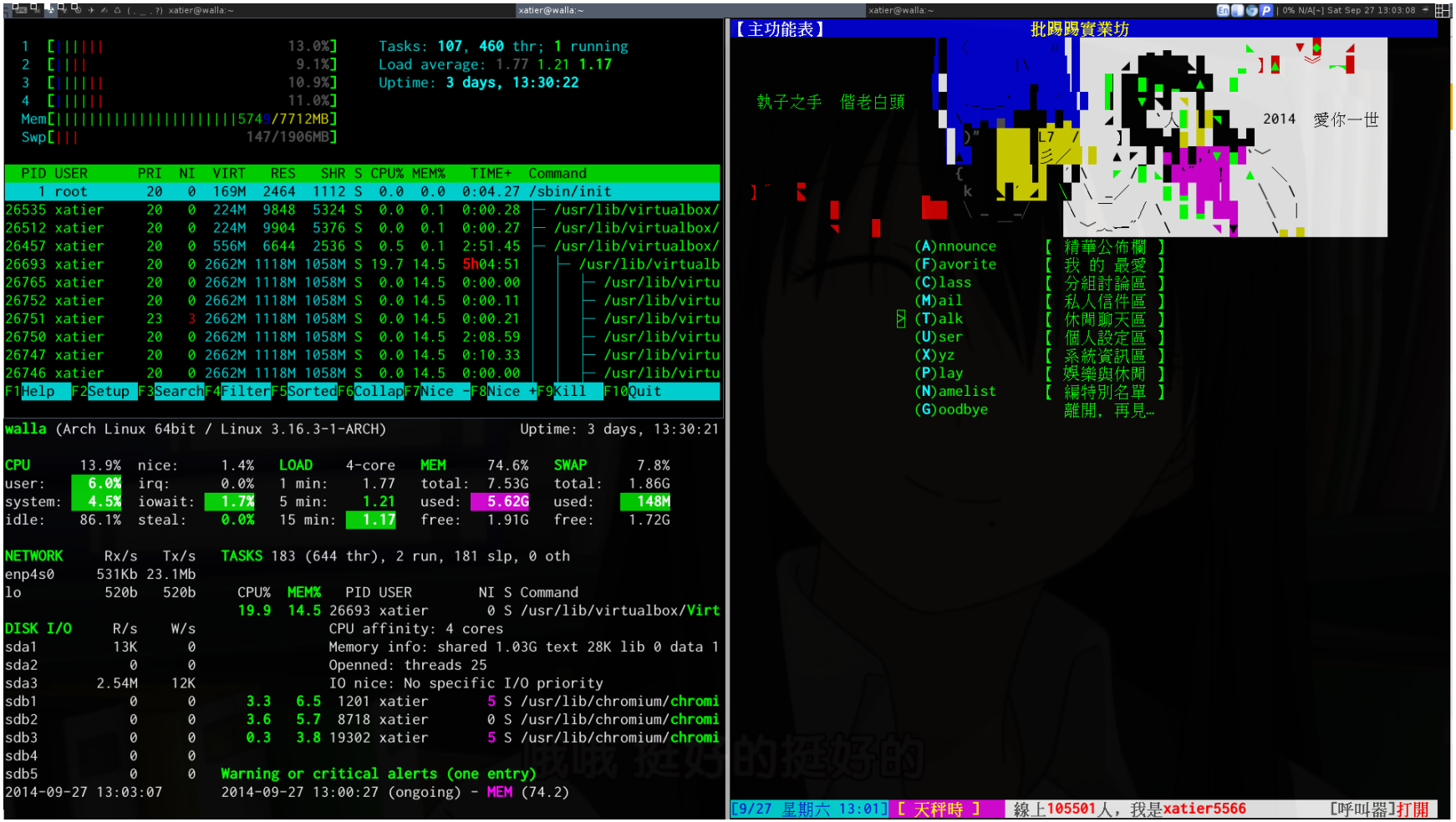
```
Section "Files"
        ModulePath   "/usr/local/lib/xorg/modules"
        FontPath     "/usr/local/lib/X11/fonts/misc/"
        FontPath     "/usr/local/lib/X11/fonts/TTF/"
        FontPath     "/usr/local/lib/X11/fonts/OTF"
        FontPath     "/usr/local/lib/X11/fonts/Type1/"
        FontPath     "/usr/local/lib/X11/fonts/100dpi/"
        FontPath     "/usr/local/lib/X11/fonts/75dpi/"
        FontPath     "/usr/local/lib/X11/fonts/TrueType/"
        FontPath     "/usr/local/lib/X11/fonts/local/"
EndSection
```

安裝中文 Terminal Emulator

- rxvt-unicode
  - /usr/ports/x11/rxvt-unicode
- ROXterm
  - /usr/ports/x11/roxterm
- mlterm
  - /usr/ports/x11/mlterm
- aterm
  - /usr/ports/chinese/aterm
- eterm
  - /usr/ports/chinese/eterm

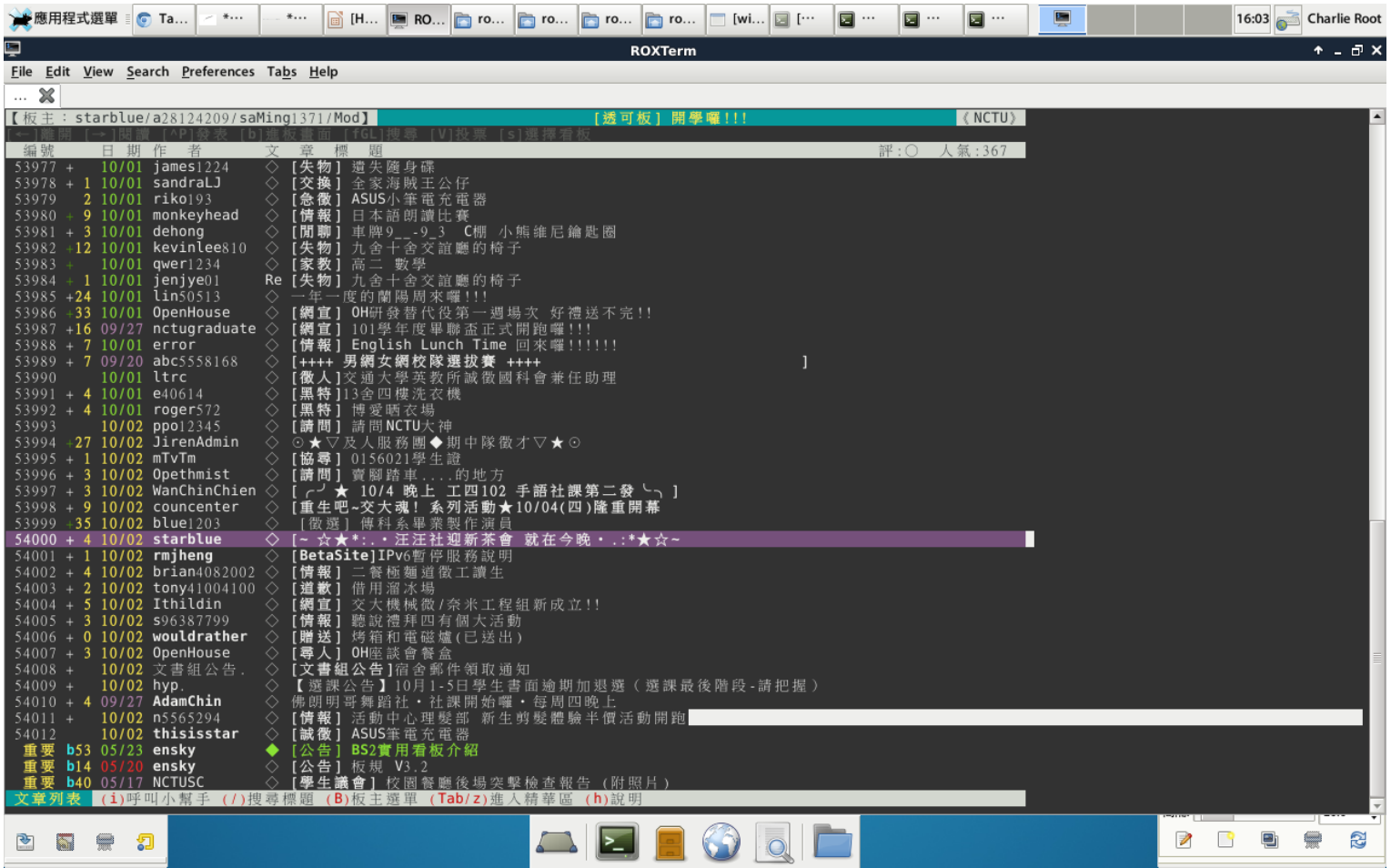
rxvt-unicode

- X11/rxvt-unicode



ROXterm

- X11/roxterm
- roxterm-config



安裝中文輸入程式

- Choices
  - ibus-chewing(chinese/ibus-chewing)
  - ibus-pinyin(chinese/ibus-pinyin)

安裝 ibus 中文輸入程式 (1)

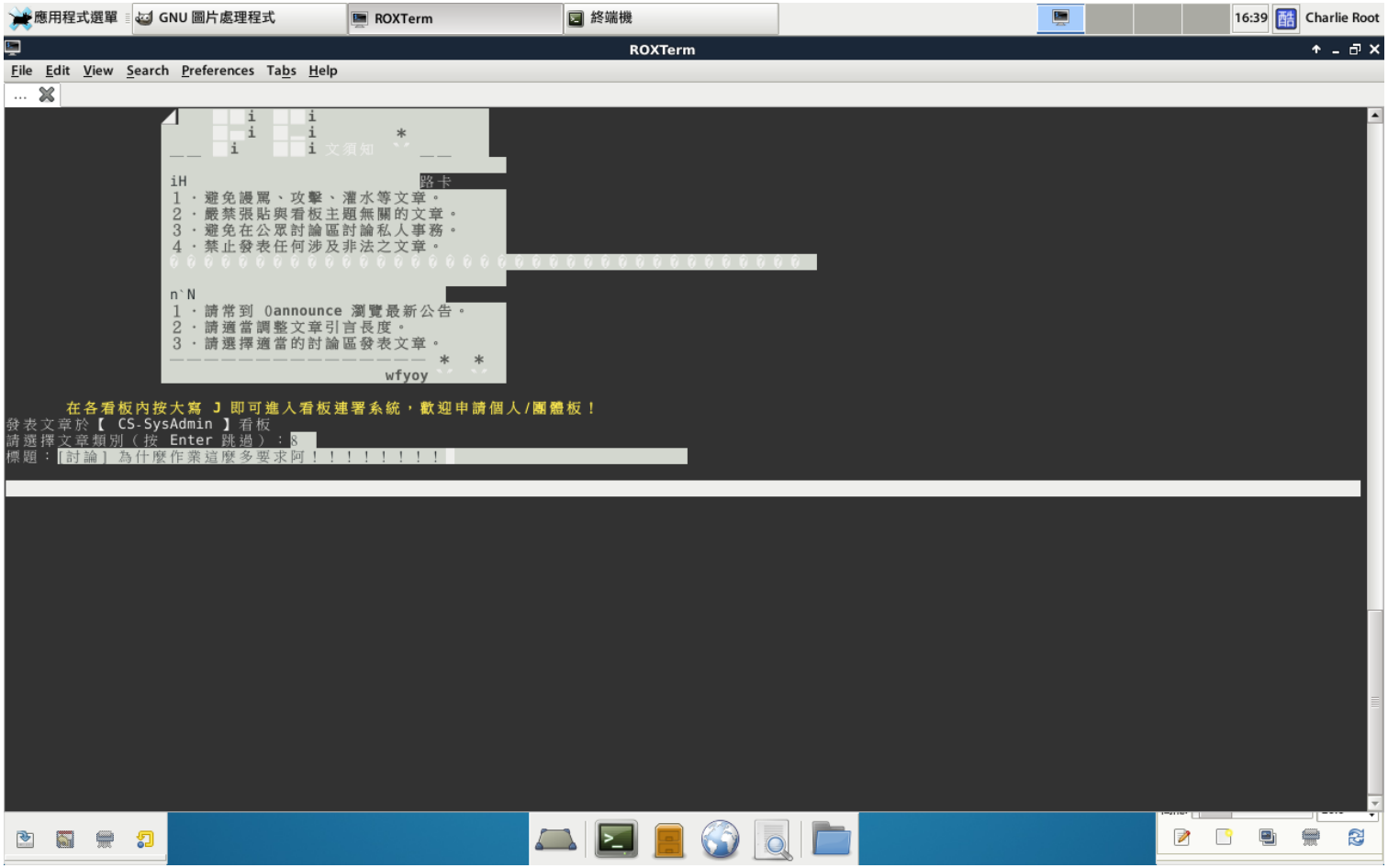
- ibus
  - Intelligent Input Bus

1. % cd /usr/ports/textproc/ibus-chewing ; make install clean
2. setenv LC_CTYPE zh_TW.UTF-8 (csh/tcsh)
   export LC_CTYPE=zh_TW.UTF-8 (sh/bash)
3. Edit .xinitrc (或是可以 setenv in .cshrc/.bashrc)

```
XIM=ibus
GTK_IM_MODULE=ibus
QT_IM_MODULE=xim
XMODIFIERS='@im=ibus'
XIM_PROGRAM="ibus-daemon"
XIM_ARGS="--daemonize --xim"
```

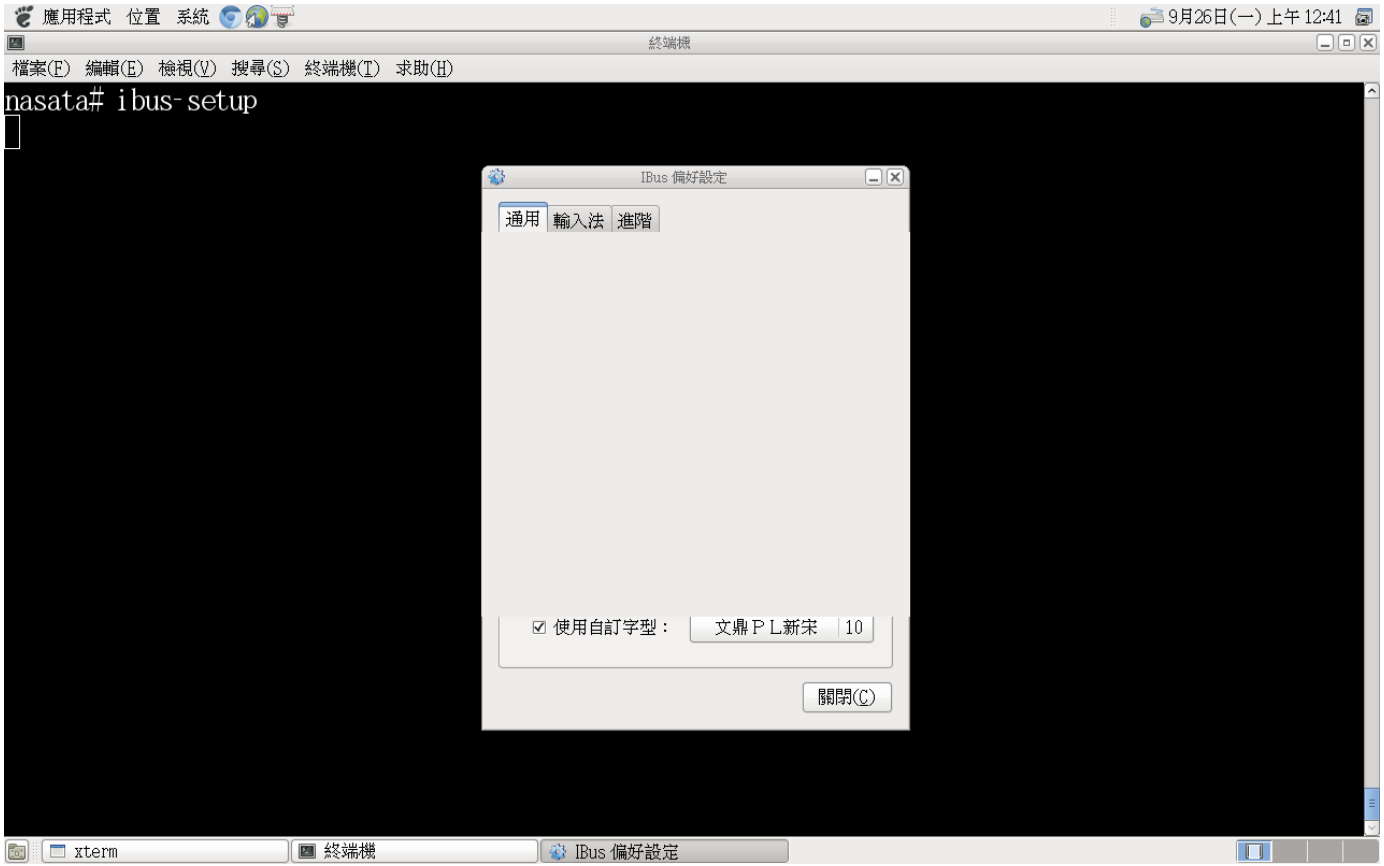
安裝 ibus中文輸入程式 (2)

4. Switch to chinese input: Ctrl -Space



安裝 ibus 中文輸入程式 (3)

- Preference
  - % ibus-setup (UTF-8)



References

- 中文碼介紹
  - http://www.cns11643.gov.tw/AIDB/encodings.do
- FreeBSD Chinese HOWTO
  - http://netlab.cse.yzu.edu.tw/~statue/freebsd/zh-tut/index.html
- Introduction to i18n
  - http://www.debian.org/doc/manuals/intro-i18n/
- Unicode 介紹
  - http://www.csie.ntu.edu.tw/~p92005/Joel/Unicode.html