Localizing GNU/Linux and XFree86

A Thailand's Experience

Theppitak Karoonboonyanan
thep@linux.thai.net

January 2004

Locale
Overview of . . .
I18N in X11R6
X Input Method
Font Systems

Localizing GNU/Linux and XFree86

- Locale
- Overview of GNU/Linux Desktop I18N
- I18N in X11R6
- X Input Methods
- Font Systems

Locale
Overview of . . .
I18N in X11R6
X Input Method
Font Systems

1. Locale

Locale
Overview of . . .
I18N in X11R6
X Input Method
Font Systems

What is Locale?

- **Internationalization (I18N):** framework abstracting culture-dependent behaviors
- **Locale:** customization according to specific cultures
- Examples:
  - **POSIX Locale:** locale for C library
  - **GNU gettext:** locale for message translation

Locale
Overview of...
I18N in X11R6
X Input Method
Font Systems

Locale Naming

OpenI18N:

$$\langle lang \rangle_\langle territory \rangle . \langle codeset \rangle [@\langle modifiers \rangle]$$

- $\langle lang \rangle$ = language code (ISO 639)
- $\langle territory \rangle$ = country code (ISO 3166)
- $\langle codeset \rangle$ = character encoding (if omitted, locale default is assumed)
- $\langle modifiers \rangle$ = locale modifiers

Examples:

- `fr_CA.ISO-8859-1`
  = French language spoken in Canada using ISO-8859-1 character set
- `th_TH.TIS-620`
  = Thai language in Thailand using TIS-620 encoding
- `th_TH`
  ≡ `th_TH.TIS-620` in general

Character Sets

- GNU C library: Unicode-based
- Character sets described as Unicode subset

Example for UTF-8:

```
...
<U0041> /x41          LATIN CAPITAL LETTER A
<U0042> /x42          LATIN CAPITAL LETTER B
<U0043> /x43          LATIN CAPITAL LETTER C
...
<U0E01> /xe0/xb8/x81 THAI CHARACTER KO KAI
<U0E02> /xe0/xb8/x82 THAI CHARACTER KHO KHAI
<U0E03> /xe0/xb8/x83 THAI CHARACTER KHO KHUAT
...
```

Character Sets

Example for TIS-620:

```
...
<U0041> /x41    LATIN CAPITAL LETTER A
<U0042> /x42    LATIN CAPITAL LETTER B
<U0043> /x43    LATIN CAPITAL LETTER C
...
<U0E01> /xa1    THAI CHARACTER KO KAI
<U0E02> /xa2    THAI CHARACTER KHO KHAI
<U0E03> /xa3    THAI CHARACTER KHO KHUAT
...
```

POSIX Cultural Conventions

- Locale for the C language library

category	description
LC_CTYPE	character classification
LC_COLLATE	string collation
LC_TIME	date and time format
LC_NUMERIC	number format
LC_MONETARY	currentcy format
LC_MESSAGES	locale messages

POSIX Cultural Conventions

Setting Locale

- C application:

```
/* <locale.h> */
/* Set and/or return the current locale. */
extern char *setlocale (int category,
                        const char *locale);
```

For example:

```
#include <locale.h>
...
const char *prev_locale;
prev_locale = setlocale(LC_ALL, "");
```

- empty-string locale means relying on environment setting:
  1. `LC_ALL` defined → use it
  2. `LC_CTYPE`, `LC_COLLATE`, ..., `LC_MESSAGES` defined → use them
  3. `LANG` defined → use it
  4. Otherwise → fall back to `C` (aka `POSIX`)

POSIX Cultural Conventions

LC_CTYPE

- C functions affected: ⟨ctype.h⟩

```
iscntrl()     isalnum()     isupper()
isgraph()     isalpha()     tolower()
isprint()     isdigit()     toupper()
isspace()     isxdigit()
ispunct()     islower()
```

- glibc: mostly relies on “i18n”

POSIX Cultural Conventions

LC_COLLATE

- C functions affected:
  - `strcoll()` = internationalized version of `strcmp()`
  - `strxfrm()` creates sorting key so that `strcmp()` yields the same order as applying `strcoll()` to source strings
- glibc: relies on ISO/IEC 14651 *Common Tailorable Template (CTT)* by default

  → localizers should check the "i18n" definition before tailoring it

  → Thai & Lao = exceptional cases
- locale definition:
  - multi-pass comparison → multi-level weight
  - ISO/IEC 14651 = 4 levels
  - describe weight as "IGNORE" to skip char

POSIX Cultural Conventions

LC_TIME

- C functions affected:
  - `strftime()`
- locale definitions:
  - day of week & month names translation
  - appropriate date & time formats
  - era

Locale
Overview of...
I18N in X11R6
X Input Method
Font Systems

POSIX Cultural Conventions

LC_NUMERIC & LC_MONETARY

- C functions affected:
  - `localeconv()` retrieves information for both categories
  - (non-standard) `strfmon()` format monetary amounts
- locale definitions:
  - LC_NUMERIC: decimal point, thousand separator, number grouping
  - LC_MONETARY: currency symbol (ISO 4217), monetary formats (debit/credit, numeric format)

POSIX Cultural Conventions

LC_MESSAGES

- C functions affected: *none*
- locale definitions: yes/no responses
- mostly used by *GNU gettext*

ISO/IEC 14652

ISO/IEC 14652 = extended POSIX, plus:

category	description
LC_PAPER	paper size
LC_NAME	personal name format
LC_ADDRESS	address codes and format
LC_TELEPHONE	telephone number
LC_MEASUREMENT	measurement units
LC_VERSIONS	locale version

- already supported by glibc
- info can be retrieved using `nl_langinfo()`
  (in fact, all locale data can be retrieved as such)

Building Locale

- Files needed
  - locale definition
  - charmap
- Symbolic names for char are usually Unicode-based
- To compile locale data:

  `localedef` [`-f` ⟨*charmap*⟩] [`-i` ⟨*input*⟩] ⟨*name*⟩

  for example:

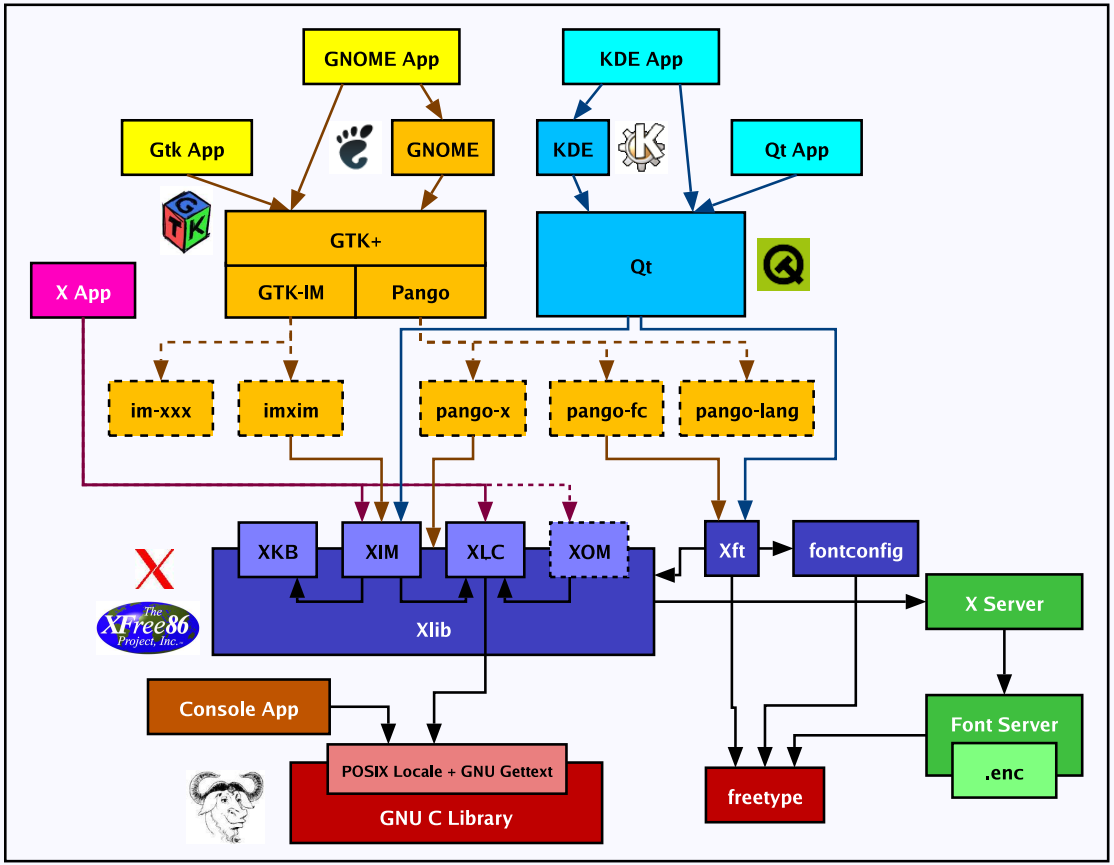
```
# localedef -f TIS-620 -i th_TH th_TH
```

- The `locale` command:
  - argument-less ⟶ list values as affected by env
  - "`-a`" ⟶ list installed locales
  - "`-m`" ⟶ list supported charmaps

2. Overview of GNU/Linux Desktop I18N

Locale
Overview of . . .
I18N in X11R6
X Input Method
Font Systems

GNU/Linux Desktop



Locale
Overview of . . .
I18N in X11R6
X Input Method
Font Systems

GNU/Linux Desktop

GNU C library

- ISO/IEC 14652 locales (POSIX proper superset)

XFree86 (X11R6.5)

- X locale (XLC)
- X Input Method (XIM)
- X Output Method (XOM)

Locale
Overview of . . .
I18N in X11R6
X Input Method
Font Systems

GNU/Linux Desktop

GTK+ 2

- Unicode-based Multilingual System
- GTK+ Input Method (gtk-immodules)
- Pango
  - rendering engine: X, Xft, OpenType backends
  - language engine: cursor movements, word break

Qt 3

- Unicode-based Multilingual System
- XIM-based input method
- Unicode-based complex text support (`QComplexText` class)

Locale
Overview of...
I18N in X11R6
X Input Method
Font Systems

3. I18N in X11R6

Locale
Overview of . . .
I18N in X11R6
X Input Method
Font Systems

I18N in X11R6

Xlib: I18N via X locales

- X locale components:
  - **XLC** – locale data
    - **XLC_FONTSET** – font charset encoding, used by XOM
    - **XLC_XLOCALE** – character encoding, conversion, etc.
  - **XIM** – input method
  - **XOM** – output method

4. X Input Method

Locale
Overview of...
I18N in X11R6
X Input Method
Font Systems

Keyboard Maps

X11R6: XKB Extension

- **XKB:** scan code $\xrightarrow{model}$ keycode $\xrightarrow{layout,variant}$ keysym
- **scan code:** raw scan code from kernel
- *model:* hardware-dependent (`pc104`, `pc105`, `microsoft`, `microsoftplus`, ...)
- **keycode:** symbolic values for *unscreened* key positions
- *layout:* language-dependent layout
- *variant:* sub-type of layout (mostly *"nodeadkeys"* for Latin)
- **keysym:** symbolic values for *screened* keyboard

Keyboard Maps

Setting up XKB

- `setxkbmap` command
  - *rules:* XFree86 4.x default method
    - normally `"xfree86"`
    - rule files: `/etc/X11/xkb/rules`
    - rules: determine *layout*, *variant*, *option* to produce the symbols map
  - *model:* hardware model
  - *layout:* up to 64 groups mixture
  - *variant:* pick up variation
  - *option:* additional configurations
    - group switching key
    - level-3 shift
    - Ctrl and CapsLock rearrangement
    - LED indicator
    - CapsLock behaviors
    - etc.

Keyboard Maps

Setting up XKB

- `setxkbmap` command (cont.)
  Example:

```
$ setxkbmap us,th \
    -option grp:alt_shift_toggle,grp_led:scroll
```

  Full options:

```
$ setxkbmap -rules xfree86 \
    -model pc104 -layout "us,th" \
    -option "grp:alt_shift_toggle,grp_led:scroll"
```

Keyboard Maps

Setting up XKB

- /etc/X11/XF86Config:

```
Section "InputDevice"
  Identifier "Generic Keyboard"
  Driver     "keyboard"
  Option     "CoreKeyboard"
  Option     "XkbRules"   "xfree86"
  Option     "XkbModel"   "pc104"
  Option     "XkbLayout"  "us,th"
  Option     "XkbOptions" "grp:alt_shift..."
EndSection
```

Providing a Keyboard Map

- New keymap → new symbols file (`/etc/X11/xkb/symbols`)
- Default symbols used by XFree86 4.3.0 default rules: `/etc/X11/xkb/symbols/pc/*` (one group per file, no pre-combined)
- Symbols file:
  - `xkb_symbols` data
  - entries:
    - group name
    - keysyms for each keycode (unshifted + shifted)
  - keysyms:
    - predefined (`<X11/keysymdef.h>`)
    - Unicode keysyms (`0x100`*xxxx*)

Providing a Keyboard Map

- Example: /etc/X11/xkb/symbols/pc/th_tis

```
partial default alphanumeric_keys
xkb_symbols "basic" {
  name[Group1]= "Thai (TIS-820.2538)";
  ...
  key <TLDE> {[0x1000e4f, 0x1000e5b]};
  key <AE01> {[Thai_baht, Thai_lakkhangyao]};
  key <AE02> {[slash,     Thai_leknung]};
  ...
};
```

- Regenerate the symbols list:

```
# cd /etc/X11/xkb/symbols
# xkbcomp -lhlpR '*' -o ../symbols.dir
```

Providing a Keyboard Map

- Extra:
  - add entry to `/etc/X11/xkb/rules/xfree86.lst`
- XFree86 source:
  - `xc/programs/xkbcomp`

Locale
Overview of...
I18N in X11R6
X Input Method
Font Systems

X Input Method

XIM

- Locale-based framework for text input of any language
- Structure:
  - X client: Input Context (XIC) for text entry
  - XIM: service (server or library)
  - X client passes key events to XIM
  - XIM translate keycode and determine action upon keysym, based on current state of XIC
  - XIM commits character back to X client event queue when conversion is ready

Locale
Overview of...
I18N in X11R6
X Input Method
Font Systems

X Input Method

User's point of view

- In general, XIM-aware app calls `XSetLocaleModifiers("")`, which allows user to choose XIM via `XMODIFIERS` env.
- For example:

```
$ export LANG=th_TH.TIS-620
$ export XMODIFIERS="@im=Strict"
```

  which specifies `Strict` input method for Thai locale.

5. Font Systems

Locale
Overview of . . .
I18N in X11R6
X Input Method
Font Systems

Traditional X Font System

X Fonts

- BDF, PCF, Type 1, TrueType
- Owned and Managed by X server or X Font Server (XFS)
- Naming: X Logical Font Description (XLFD)

```
-adobe-times-medium-r-normal--*-100-75-75-p-
54-iso8859-1
```

Traditional X Font System

X Fonts

- Installing:
  - Prepare `fonts.dir` with `mkfontscale` & `mkfontdir`
  - Add the font path to either:
    - X server (immediate):

```
$ xset fp+ /your/font/path
$ xset fp rehash
```

    - X server (next startup): /etc/X11/XF86Config:

```
Section "Files"
  ...
  FontPath "/your/font/path"
  ...
EndSection
```

    - XFS: /etc/X11/fs/config:

```
catalog=...,/your/font/path
```

      then:

```
# /etc/init.d/xfs restart
```

Xft and Fontconfig

XFree86 4 introduces:

- **X Render Extension:** alpha, anti-aliasing, sub-pixel (server side)
- **Xft Library:** rasterize w/ FreeType & X Render (client side)
- **Fontconfig Library:** font accessing (client side)

→ Client-side fonts + Anti-aliased

Xft/fontconfig

- Used by GNOME 2 & KDE 3
  → becomes default usage
- Traditional X fonts still required by GTK+ 1.2 apps, though

Xft and Fontconfig

Configuration

- `/etc/fonts/fonts.conf`
  - XML format
  - `<dir>` lists font (root) directories
  - others: instruction for font matching & customization
- Common default config:
  - System fonts at `/usr/share/fonts/`
  - User fonts at `~/.fonts/`
- Installing fonts:
  - Copy fonts into one of the defined font paths
  - Regenerate font cache

```
# fc-cache -f
```

- Listing fonts:

```
# fc-list
```

Locale
Overview of...
I18N in X11R6
X Input Method
Font Systems