THE ARCHITECTURE OF
CHINESE-ENGLISH MICROCOMPUTER
SYSTEMS
by
S.J. Huang and R.T. Liu
Nankai University, Tianjin,
People's Republic Of China
and
N.P. Archer and M.W.L. Chan
Faculty of Business
McMaster University
Hamilton, Ontario
Canada L8S 4M4
December 1986
Working Paper #266
THE ARCHITECTURE OF
CHINESE-ENGLISH MICROCOMPUTER
SYSTEMS
by
S.J. Huang and R.T. Liu
Nankai University, Tianjin,
People's Republic Of China
and
N.P. Archer and M.W.L. Chan
Faculty of Business
McMaster University
Hamilton, Ontario
Canada L8S 4M4
December 1986
Working Paper #266
THE ARCHITECTURE OF CHINESE-ENGLISH MICROCOMPUTER SYSTEMS

by

S.J. Huang and R.T. Liu
Nankai University, Tianjin, People's Republic Of China

and

N.P. Archer and M.W.L. Chan
Faculty of Business
McMaster University, Hamilton, Ontario
Canada L8S 4M4

December 1986
THE ARCHITECTURE OF CHINESE-ENGLISH MICROCOMPUTER SYSTEMS

ABSTRACT

Computing in the People's Republic of China has recently begun to gain momentum and, because of China's large population, its computing requirements are of worldwide interest. Much of the recent advances in computing in the People's Republic of China (PRC) have occurred in the area of microcomputer software. In particular, systems have been developed which allow the use at will of both Chinese and English as needed. This greatly improves the ease of using already developed English application software, but with a Chinese language interface. The greatest difficulty in developing Chinese language software has been in the entry of Chinese characters, but a number of efficient methods have been developed to handle this problem. Methods of displaying, printing, and communicating Chinese characters, as well as mixing Chinese with English on microcomputer systems, have all been developed and are constantly being improved.
THE ARCHITECTURE OF CHINESE-ENGLISH MICROCOMPUTER SYSTEMS

INTRODUCTION

The use of computing in the People's Republic of China (PRC) is increasing, but a major impediment to the widespread use of computers has been the adaptation to computing of written Chinese, which has thousands of complex characters. There have been several projects in Western countries which have addressed this problem [5,6,13,15,18,21], but there have also been recent developments in Taiwan [2,12]. A great deal of progress has also occurred in Japanese text processing [14], and computer handling of Japanese text is related to Chinese since the Kanji characters used in Japanese are of Chinese origin. In addition, there have been at least 400 projects [23] undertaken on the handling of Chinese character computer input, by universities and other organizations in the PRC (see for example, [19]). An indication of the widespread interest and advancement in Chinese computing can be gained by examining the Proceedings of the 1986 International Conference On Chinese Computing [16] which contains no less than 70 papers (in English) on the subject.

Part of the current growth in computer use in the PRC is due to the increased use of mainframes and minicomputers [8,22], with associated development in Chinese character systems [1,10,20,23]. There is still a shortage of computing facilities in the PRC [4]. However, a large part of the recent growth in the use of computing in the PRC has been with low cost microcomputers, and much progress has been made recently with Chinese-English
microcomputer systems. These systems allow both Chinese and English to be used, thus gaining the advantage of access to existing English software and systems.

Conversion of an English-based ASCII microcomputer system is possible by inserting a software interface between the operating system and the input/output control system so that Chinese characters are detected and managed in a special way. With the appropriate system software, the modified system can then support programs for input, display, print and communication of both types of characters, and the user can input either Chinese or English characters on the same keyboard, display them on the same screen, and print them on the same printer. When the system software has been modified to handle Chinese as well as ASCII characters, then much existing applications software can be modified to include Chinese characters where text is used, such as in menus, etc. This allows the Chinese user access to the software as easily as the English user for which it was originally designed.

Chinese-English system packages have been developed for some of the more popular makes of microcomputers such as the IBM PC, Apple II, and the Tandy TRS-80. Packages now being used include the Dragon system, developed in Taiwan [2,3], the NK system, developed at Nankai University, and a system developed by the Electronic Research Institute. The latter two institutions are in the People's Republic of China. The examples given in this paper will be based on the implementation of the NK system, because of our familiarity with it. Figure 1 shows the initial screen for the NK-DOS system.
"Welcome To The Nankai University Chinese Computer System.
Do You Want To Use Chinese? (Y/N)"

Opening Screen From NK-DOS Microcomputer System

CODING SYSTEMS FOR CHINESE CHARACTERS

Many Chinese words consist of only one character, but there are many more with two, three and sometimes four or five characters. As well, the average number of characters in a sample of English text is about 3.7 times that of Chinese characters in equivalent text [7]. Since Chinese is an ideographic language there are many thousands of Chinese characters, but the Chinese character set is relatively uniform throughout the PRC, even though there are eight major distinct languages spoken in the PRC
and many more dialects within these languages [9]. Mandarin Chinese (Putonghua) is the standard language in the PRC, which provides a standardizing influence used to advantage in the class of computer input methods which rely on phonetic computer input techniques.

Any one of the many techniques which have been developed may be used to input Chinese characters through a keyboard. While there are many ways of coding the Chinese characters for input, the internal representation of each character in the computer should adhere to some standard. Unfortunately, there is no single standard for internal coding of Chinese characters. Some coding systems use two bytes, but others use three and some use six byte codes for internal representation of each character. The most popular coding system in the PRC uses the two byte internal code which is faster to manipulate. The disadvantage is that the code must be transformed before being communicated to other systems since all of the 16 bits in the two byte code are used to represent the character, and information may be lost during transmission (with the two byte code, the most significant bit (MSB) of each byte is flagged as a "1" if it is a Chinese code and a "0" if it is an ASCII code). Japan has also adopted a two byte standard code JIS C6226 to represent Kanji characters in a similar manner, using seven bits of each byte [17].

Standards developed thus far for Chinese character codes also do not necessarily cross national boundaries and this inhibits the development of data communication systems. Modern reform of the Chinese character set in the PRC by eliminating
variant forms and simplifying characters now requires translation of Chinese characters between the PRC and other countries using the old Chinese character set. However, while traditional Chinese writing is in vertical columns, modern Chinese usage in business and technical writing is from left to right, which simplifies adaptation of the language to Western computer systems.

The PRC has also recently defined a Standard Chinese Character Code for Information Interchange (GB 2312-80) which contains two byte codes for the 6763 most commonly used Chinese characters in two classes; the first contains 3755 and the second most frequently used class contains 3008. It also includes codes for characters from other languages such as English and Russian. At the same time, for example, Taiwan has developed a different Chinese character code [11] for information interchange (CCCII) which contains 4807 of the most frequently used and 16,197 less frequently used characters. There is also an International Syllabus of Chinese Characters with another coding scheme.

**CODING FOR INPUT**

The most difficult problem facing the designer of a Chinese language computer system is in coding characters for efficient and easy input. Once the characters have been input, the problem of displaying, printing or transmitting Chinese characters can be handled as long as there is an agreed-upon code for the internal coding of the Chinese characters. Input coding techniques for Chinese characters may be separated into four general groups. These groups include coding by whole character, coding by sound,
coding by character shape, and coding by mixing sound and shape.

Coding by whole character is based upon the use of a standard numerical code for the input of each Chinese character. There are two well-known standards for these unique codes. The four digit telegraph code is one of the earliest systems and it is still in use. A second system is the GB2312-80 standard character set which can be represented by a two byte code. The obvious difficulty with using either of these systems is that it is impossible to memorize the thousands of codes needed to represent the Chinese characters, and users must resort to looking the codes up in a table. Other methods use the whole character code through special keyboards. A big keyboard which has been used contains up to 4000 keys, one for each of the more commonly used Chinese characters. But searching a keyboard this size to find the required key is tedious and awkward. A "middle" keyboard has many fewer keys, but allows the use of a special key to select from among many possible "pages" of character images on the same keyboard. While this is an improvement on the big keyboard, it still requires a great deal of practice to become proficient in its use.

A second technique, coding by spelling, uses the phonetic symbols of Mandarin Chinese. These are written in Roman characters, a phonetic spelling known as "hanyu-pinyin". The difficulty with the phonetic entry of Mandarin characters is that there may be many characters with the same pronunciation (homonyms). To solve this problem, the homonym set for each phonetic character may be displayed automatically on the screen when it is entered, allowing the user to select the character.
actually desired.

A third method of coding for character entry involves coding by the shape or position of the character. There are many ways of breaking down character shapes into various types of fundamental characteristics. These approaches are feasible, but very tedious for the more complex characters. A method of building up some of the more complex characters from duplications of simple characters can also be used. For example, the Chinese character for "wood" which is shown in Figure 3 in the following text can be repeated to form a new character which is a combination of two of the original characters shown, and this is the Chinese character meaning "forest". Again, this technique is complex because of the knowledge required about character structure.

There are additional techniques which attempt to combine the features of shape and sound of Chinese characters, and several more advanced techniques. For example, Becker [6] describes a multilingual word processing system developed by Xerox which uses phonetic entry of entire Chinese words. The occurrence of homonyms in words is far less likely than with characters, and therefore users of this technique are not often required to make choices among homonyms. The difficulty with this approach is the need to store far more information in the form of a Chinese dictionary which is searched by the computer for each word entered, a time-consuming process compared to searching a more limited set of characters which may be stored in main memory. The Tian Ma system described by Naegele [15] is contained within
an add-on board for a microcomputer and uses automatic context search to select the appropriate homonym. This system is claimed to have a 99 percent "hit" rate for the correct character.

While coding by phonetic spelling of characters or words allows input at a reasonable rate of speed, users who are not very familiar with Mandarin Chinese and its phonetic spelling will still have difficulty in entering text. A period of training is required to use this technique, but most Chinese children are trained in phonetic Mandarin Chinese spelling, providing an educational background to build upon.

Because users may wish to use more than one input technique, depending upon the knowledge of the user or the particular text being input, it is customary to offer several input options to the user of a Chinese system. To illustrate, Figure 2 shows the input option screen for the NK-DOS system.

**INPUT SYSTEMS**

If both Chinese and English characters are to be handled by a computer system, there will be four different internal codes:

- ASCII (including the control characters)
- Chinese character input
- Chinese character internal
- Dot matrix Chinese display

The projects to investigate Chinese character computing have resulted in many different techniques for processing input characters. Any software for handling the input of Chinese characters may allow several such options for character input, requiring software to handle each such option. However, all such
character information must be transformed into a standard internal character code set which can be processed by the character management software. In the newer microcomputer-based systems, keyboard control is through operating system functions. For example, in NK-DOS an expanded 1 KByte of RAM is used for the keyboard buffer. The input from the keyboard is stored in a FIFO queue. NK-DOS analyzes the keyboard input to determine whether it is a control code, Chinese character input code or ASCII code.

Figure 2.

Input Technique Selection For NK-DOS Microcomputer System
If it is a control code, a special subroutine is called to analyze and perform the required function. The entire keyboard interface program occupies roughly 10 KBytes of memory space and is stored in EPROM.

**CHINESE CHARACTER PRINTING**

Chinese characters are usually printed on dot matrix or laser printers. Because of the complexity of Chinese characters, a resolution of at least 16x16 pixels is required to display the characters on printers or screens (see Figure 3 for an example). For most printers, the Chinese character sets are not included in ROM, so they must be generated by the computer and transmitted in graphics form to the printer. The entire character set must be stored in the computer database in a form that can be suited to the resolution of the printer. Chinese characters are double the height of ASCII characters, and the print program must be able to separate the Chinese characters (16x16, 24x24 or 32x32) into component rows so a line of characters may be printed in several passes if needed to obtain the desired resolution.

For printer control under NK-DOS, for example, the Chinese print function calls must cover a much wider spectrum of functions than the English function counterparts. Depending upon the type of printer that is used, NK-DOS will execute different print programs. The major function of the print program is to create a Chinese print buffer so that one row of Chinese characters in dot matrix code will be stored and eventually sent
to the printer in graphic print mode. The process includes acquiring printer information, setting up and converting the dot matrix for each character, and deciding on print format. In printing the Chinese characters with different sizes, NK-DOS relies on the MS-DOS functions for execution.

**CHARACTER DISPLAY**

Chinese characters can be displayed on standard screens in 16x16 pixel cells but higher resolution systems could use 24x24, 32x32 or even more pixels per character. Since at least one column in each character must be used as a separator, only 16x15, 24x22, or 32x30 pixels are normally used for the character itself. The display of Chinese characters is no different than for ASCII characters, but using a standard microcomputer system requires that the Chinese characters be bit mapped as special graphical characters. Figure 3 shows the bit map for the Chinese character "wood". The internal hexadecimal representation of the 16x16 bit map of the character shown in the figure is 00, 80, 00, 80, 00, 80, 80, 00, 7F, FF, 01, CO, 02, A0, 04, 90, 08, 88, 10, 84, 20, 82, 40, 81, 00, 80, 00, 80, 00, 80, 08, 00, 80.

Normally there are 25 rows and 80 columns on a standard ASCII display screen, giving 2000 characters per screen. On the IBM PC, the monochrome display system has a character generator and a 4 Kbyte memory buffer (1 byte for each character and 1 byte for its attributes) to store the codes for the characters to be displayed. If the character generator is not used, as with the "high resolution" monochrome graphics mode of the IBM color graphics adapter used by the NK-02 system, then 16 Kbytes of
display memory are available to store the necessary character bit maps for the 640x200 screen. For 16x16 characters the display memory can store only 480 character cells, and display 12 rows with 40 characters per row. For the NK-04 and later versions of the NK system, a screen resolution of 640x400 is used with a monochrome graphics adapter, allowing the display of 25 lines with 40 characters per line.

The 16 Kbytes of display memory in the NK-02 system is used for direct storage of the Chinese and ASCII character bit maps. In a Chinese-English system, an ASCII character normally takes up half the number of columns and therefore half the display space of a Chinese character.
For the screen display of text, a "display buffer" corresponds to the block of text in RAM memory which is also concurrently stored in the display memory on the graphics controller card. The normal PC display capacity for ASCII characters is 2 KBytes. Every byte in the display buffer is linked to a character in the display memory. With Chinese characters, there are usually two bytes linked to a display memory character. If the system were to display 25x40 Chinese characters on the screen, the screen display to buffer proportion would be two to one. However, a display memory that does not have exactly this capacity can lead to difficulties in designing general high level software for mixed Chinese and English character display. One way around this difficulty is to use the same buffer for mapping both Chinese and English characters. Part of it can be used as a window which maps the Chinese character display memory. The high level software recognizes this memory as a virtual display buffer and accesses it, linking the virtual display buffer to the physical display screen by moving the window about on it. There is then no need to distinguish between Chinese characters and other characters which can also be represented in dot matrix form. The standard English characters can be written into display memory by a function in the IBM BIOS (Basic InputOutput System) routines, invoked through a standard interrupt feature of MS-DOS.
To edit an ASCII screen display when a 2 Kbyte display buffer is used, one needs only to modify the display buffer and then call the BIOS display subroutine. However, in generating a graphic display, the character dot matrix as supplied in a user table is written into the 16 Kbyte display memory when the display function routine is called. Any ASCII character can also be displayed in the same manner in graphics mode. Editing of the screen (insert, delete, etc.), requires rewriting the entire display memory. If we want to use the 2 Kbyte character generator memory map which stores the ASCII character information, we can use a simple subroutine to align the display buffer with the 16 Kbyte display memory. This will allow us to edit the display memory as if we were using the character generator.

To display Chinese characters, we add a "Chinese character function call" subroutine. This is equivalent to the function call used to manage the display of ASCII characters. This routine links the Chinese part of the display buffer to the display memory. In the Nankai University NK-02 system, for example, the Chinese character area occupies 960 bytes (2 bytes per character referenced) of the total of 2 Kbytes of the display buffer. To edit this Chinese character data, it is essential that the high level editing software be able to distinguish whether the code in the buffer is ASCII or Chinese character mode. To handle this situation, the NK system uses part of the 2 KByte ASCII display buffer as the Chinese display buffer. In
conjunction with the Chinese display function routine, this is used to reference Chinese characters in the screen display memory. Since the Chinese display buffer occupies only 960 bytes of memory, the remaining space in the 2 Kbyte buffer can be used for the storage of other character references, even though the information is not displayed on the screen. By changing the low address of the 960 byte Chinese virtual display buffer, this window can be moved about in the 2K byte buffer, and thus scroll any of the characters referenced in the buffer onto the screen. The window movement is controlled by a Chinese display function call. For any high level software, this becomes a transparent interface controlled by either the ASCII or the Chinese display function calls, as determined by the initial system boot.

Other functions which are attended to by the NK-DOS software include adjustments to the cursor position, the end of row position, calculation of the character number in the row, changing the first address of the Chinese display memory map, row scroll and page turning. For communicating with the scan control chip in the graphics controller board, the original MS-DOS function calls are used, under MS-DOS control. A particular problem occurs when editing characters at the end of a line in a Chinese-English system. This is due to the fact that the Chinese characters are twice the width of the ASCII characters so that a combination of Chinese and ASCII characters may not fit the line length exactly. A character editing program which attempts to handle this problem can cause difficulties, especially when editing computer programs, since errors in placement on a statement line may affect program execution.
To provide a system which works easily with both Chinese and ASCII characters, the operating system must be modified so it can distinguish the Chinese characters and support the Chinese character processing routines. This is not difficult to do with hierarchical operating systems such as CP/M and MS-DOS. For example, in the IBM PC, the BIOS has many function calls which control the operation of the display terminal, keyboard, printer and disk drives. The MS-DOS operating system has an interface level IBMBIO which executes a DOS command by calling BIOS. As shown in Figure 4, if a Chinese character processing interface level program is inserted between BIOS and IBMBIO by modifying the addresses in certain standard DOS interrupts, then basic Chinese character processing routines can be used to intercept calls which involve Chinese characters for input, display, printing, or communication. Using these procedures when DOS calls BIOS, the Chinese interface level passes control to the appropriate Chinese processing function if Chinese functions are involved. If not, the call is passed directly to BIOS, as in handling ASCII characters.

The Chinese-English interface for the NK-DOS system between DOS and BIOS on an IBM PC consists of two parts. One part distinguishes between the ASCII code and the internal Chinese character code. The other part of the interface is used for controlling peripheral devices. When the machine is booted, the user is required to select either ASCII or Chinese mode. If the Chinese mode is chosen, the system automatically begins execution
of NK-DOS. Otherwise, the system operates under MS-DOS control.

---

High level software

---

DOS

---

Chinese software interface

---

BIOS

---

Hardware

---

Figure 4.
System Software Interfaces

DATA COMMUNICATIONS WITH CHINESE CHARACTERS

Data communications are still in their infancy in China [8] and there is no standard technique for asynchronous communication. As an example, the NK-02 and NK-03 systems can transmit on both Omninet and Ethernet LANs (local area networks), but neither system is ready for commercial network application at this time.

Some of the difficulty with Chinese character communication is due to the complexity of handling the large character set. When a two byte internal code is used for Chinese characters, all 16 bits are used for data. For synchronous communication, all 16
bits are transmitted. However, for most asynchronous communication systems the MSB (most significant bit) in each byte is used for parity and this may destroy Chinese character information being transmitted. Therefore, before characters are transmitted, it is necessary to transform the data in order to avoid the loss of this identifying information. This is often done by using more than two bytes for the Chinese character information. The data transmitted may then be reassembled into the original two byte code at the destination. Some Chinese character systems add a header in front of a Chinese character string. This header may be only one byte, but it is often two to four bytes. For a four byte header, the first two bytes will be a specific flag to identify the following string as Chinese characters. The next two bytes in the header will contain such information as the packet size in bytes, including the header. A packet with a four byte header can contain up to 8 KBytes. Each Chinese character will use two bytes including the MSB, even though the MSB could be masked during communication to another system. If the Chinese system interface program is processing information for communication when it includes a Chinese character string, it puts headers in front of strings to be transmitted. It also strips headers off the strings it receives and re-sets MSBs to 1 for Chinese characters.

**DATABASE OF CHINESE CHARACTERS**

All together there are about 7500 characters in common use in the People's Republic of China, including the 6763 first and second most commonly used groups of Chinese characters, and counting
Roman, Greek, Japanese kana, and special characters. In creating a database of characters, space should be reserved for about 8000 characters. For a 16x16 dot matrix, the database would need 256 Kbytes of memory. For a 24x24 dot matrix representation of each character, about 576 Kbytes is needed. The database may be stored on floppy disk, hard disk, RAM, ROM or EPROM. While the floppy disk may have enough storage for the 16x16 dot matrix format, it is too slow. A hard disk may also not be fast enough for some applications. The entire database of dot matrix codes may also be transferred into RAM when the program is being booted. However, this takes up so much RAM storage that it may limit the application programs which may be used, particularly on most current PCs which are limited to 640 Kbytes of RAM storage. The best option is to store the Chinese character database in ROM or EPROM on a card which is inserted into the microcomputer system. This database can be shared by the printer and the screen. Some printers allow the installation of a Chinese ROM in the printer, thus simplifying the printing of the combined Chinese and ASCII character sets.

While most commercially available systems supply the basic set of Chinese characters for input, display and printing purposes, it may be necessary to develop special characters which are not available with the system. For this, a program must be provided for the user to develop a dot matrix display map of the character and store it in an auxiliary database of characters, along with the storage code which will be used to manipulate the character internally.
ACKNOWLEDGMENTS

This work was supported partly by the Chinese State Education Commission and the World Bank through financial assistance for Professor Huang and Professor Liu during a one year visit at McMaster University, and partly by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES


<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Author(s)</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>247</td>
<td>Approximately the Statistical Properties of Elasticities Derived from Translog and Generalized</td>
<td>Itzhak Krinsky and A. Leslie Robb</td>
<td>February 1986</td>
</tr>
<tr>
<td></td>
<td>Leontief Cost Functions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>248</td>
<td>An Integration of Kelley's Covariation and Configuration Theories</td>
<td>John W. Medcof</td>
<td>March 1986</td>
</tr>
<tr>
<td>249</td>
<td>The International Labour Organization's Standards and the African Labour Laws</td>
<td>Isik Urla Zeytinoglu</td>
<td>March 1986</td>
</tr>
<tr>
<td>251</td>
<td>The Role of Management in a Political Conception of The Employment Relationship</td>
<td>Roy J. Adams</td>
<td>June 1986</td>
</tr>
<tr>
<td>252</td>
<td>Employment Standards As Industrial Relations Sub-System: The Ontario Case</td>
<td>Roy J. Adams</td>
<td>June 1986</td>
</tr>
<tr>
<td>253</td>
<td>Recruitment and Selection of Visible Minorities in Canadian Police Forces: A Survey of Selected Police Agencies</td>
<td>Harish C. Jain</td>
<td>June, 1986</td>
</tr>
<tr>
<td>254</td>
<td>An Algorithm To Generate the Ideals of a Partial Order</td>
<td>George Steiner</td>
<td>July 1986</td>
</tr>
<tr>
<td>255</td>
<td>A Theoretical Basis For Scheduling Mixed-Model Assembly Lines For Just-In-Time Production Systems</td>
<td>John Miltenburg</td>
<td>July 1986</td>
</tr>
<tr>
<td>256</td>
<td>Scheduling Mixed Model Multi-Level Just-In-Time Production Systems</td>
<td>John Miltenburg</td>
<td>July 1986</td>
</tr>
<tr>
<td>258</td>
<td>The Effects of Event Valence and Degree of Attribution Upon Action Recommendations</td>
<td>John W. Medcof</td>
<td>September 1986</td>
</tr>
<tr>
<td>259</td>
<td>One-Dimensional Floor Layout Solutions Using Squared Distances</td>
<td>Paul D. Dowling and Robert F. Love</td>
<td>September, 1986</td>
</tr>
<tr>
<td>261</td>
<td>Catalyzing Interfunctional Efforts to Find and Creatively Solve Important Business Problems</td>
<td>Min Basadur</td>
<td>September, 1986</td>
</tr>
</tbody>
</table>


