```
                    ########
               ##################
          ######              ######
       #####
      #####   ####  ####     ##      #####   ####  ####  ####  ####  ####   #####
     #####    ##    ##     ####    ##  ##   ##  ##  ###      ##    ####  ##    ##    ##
    #####    ########    ##  ##   ##        #####        ##      ## ## ##     ##
    #####    ##    ##   ######## ##  ##   ##  ###      ##      ##  ####    ##    ##
    #####   ####  ####  ####  #### #####   ####  ####  ####  ####  ####   ######
    #####                                                                      ##
     ######            ######           Issue #20
      ##################              April 18, 2001
         ########
```

....

        Ineptitude: If you can't learn to do something well, learn to enjoy
                doing it poorly.
                                        -- "Demotivator" poster

....

BSOUT

Hoo-ah! Time for another late (but hopefully great) issue of C=Hacking!

This issue features several nifty articles on both software and hardware
projects. There seems to be a lot of activity in the hardware area right
now, so hopefully we'll see more hardware articles in future issues.

In the software area, however...

If I may pithily berate for a moment, I'd like to observe that talking about
programming on comp.sys.cbm is -- this may come as a surprise to some -- not
the same as actually programming. I'd like to once again encourage those who
have been talking about projects, or have half-completed projects sitting
around on an FD disk somewhere, to go for it and get the job done!

Just like... um... C=Hacking (hey, it gets done... eventually...).

The format for this issue has changed a little, with all the news and stuff
moved into the previously skimpy "Jiffies" section. Therefore, this is now
just the 'me' column.

So 'me' would like to thank all the authors in this issue for their time
and effort (and patience), and all of the true C= Hackers out there for their
spirit and work and cool projects.

'Me' is also very happy to announce that he is getting married on August 17.
And heck, you're ALL INVITED (I think an SX-64 would make an excellent
wedding present, don't you?).

And finally, me still thinks the 64 is the niftiest computer ever made.

Onwards!

-Steve

....
....
..
.   C=H #20

:::::::::::::::::::::::::::::::::: Contents :::::::::::::::::::::::::::::::::::::

BSOUT
        o Voluminous ruminations from your unfettered editor.

Jiffies
        o News, things, and stuff.

Side Hacking

        o "Super/Turbo CPU VDC Hack", by Henry Sopko <henry.sopko@hwcn.org>
          Normally it is not possible to access the VDC chip when using
          a SuperCPU64 or Turbo CPU on a 128. The 1-wire hack described

in this article fixes that situation!

    o "16K Memory Expansion for the VIC-20", by Per Olofsson
      <MagerValp@cling.gu.se>. This article describes a nifty way to
      add more memory to the VIC-20, along with some basic circuit
      design information for the hardware neophyte (i.e. people like me!).

    o "Quick Quadratic Splines", by moi <sjudd@ffd2.com>
      A spline is a powerful tool for drawing a curve through an arbitrary
      set of points -- for motion/animation, for arbitrary curves (like
      fonts), and numerous other tasks. This article describes _quadratic_
      splines and some fast C64 implementations, and includes a program
      for experimenting with splines.

Main Articles

    o "VIC-20 Kernal ROM Disassembly Project", by Richard Cini
      <rcini@email.msn.com>

      The ever-dependable Richard Cini has written the fourth article
      in the quest for a complete disassembly of the VIC-20 kernal. This
      installment focuses on device I/O routines: SETNAM, SETLF, OPEN,
      and beyond.

    o "MODs and Digital Mixing", by Jolse Maginnis
                              <jmaginni@postoffice.utas.edu.au>.

      Josmod is a program for JOS, by Jolse, that can play Amiga MOD files
      (and their newer successors). This article describes the general
      functioning of the program, the layout of a MOD file, and how to mix
      multiple digital samples in real-time (and hence play MODs!).

    o "The C64 Digi", by Robin Harbron <macbeth@psw.ca>, Levente
      Harsfalvi <levente@terrasoft.hu>, and Stephen Judd <sjudd@ffd2.com>

      This article is, we hope, a complete reference on digital sampling
      and the C64, including: general theory, SID hardware description,
      and methods of playback (changing $d418, pulse width modulation,
      and various tricks). Numerous code examples are given, along with
      a program that does true 8-bit playback at 16KHz -- it requires a
      SuperCPU, but it is most impressive, and chances are awfully good
      that you've never heard a digi like this out of SID before.

.... Credits ....

Editor, The Big Kahuna, The Car'a'carn..... Stephen L. Judd
C=Hacking logo by......................... Mark Lawrence

Special thanks to the folks who have helped out with reviewing and such,
to the article authors for being patient, and to all the cbm hackers that
make this possible!

Legal disclaimer:
        1) If you screw it up it's your own fault!
        2) If you use someone's stuff without permission you're a dork!

About the authors:

Jolse Maginnis is a 20 year old programmer and web page designer,
currently taking a break from CS studies. He first came into contact
with the C64 at just five or six years of age, when his parents brought
home their "work" computer. He started out playing games, then moved on
to BASIC, and then on to ML. He always wanted to be a demo coder, and in
1994 met up with a coder at a user's group meeting, and has since worked
on a variety of projects from NTSC fixing to writing demo pages and intros
and even a music collection. JOS is taking up all his C64 time and he
is otherwise playing/watching sports, out with his girlfriend, or at a
movie or concert somewhere. He'd just like to say that "everyone MUST
buy a SuperCPU, it's the way of the future" and that if he can afford
one, anyone can!

Richard Cini is a 31 year old vice president of Congress Financial
Corporation, and first became involved with Commodore 8-bits in 1981, when
his parents bought him a VIC-20 as a birthday present. Mostly he used it
for general BASIC programming, with some ML later on, for projects such as
controlling the lawn sprinkler system, and for a text-to-speech synthesyzer.
All his CBM stuff is packed up right now, along with his other "classic"
computers, including a PDP11/34 and a KIM-1. In addition to collecting
old computers Richard enjoys gardening, golf, and recently has gotten

interested in robotics. As to the C= community, he feels that it
is unique in being fiercely loyal without being evangelical, unlike
some other communities, while being extremely creative in making the
best use out of the 64.

Robin Harbron is a 28 year old internet tech support at a local
independent phone company. He first got involved with C= 8-bits
in 1980, playing with school PETs, and in 1983 his parents convinced
him to spend the extra money on a C64 instead of getting a VIC-20.
Like most of us he played a lot of games, typed in games out of
magazines, and tried to write his own games. Now he writes demos,
dabbles with Internet stuff, writes C= magazine articles, and, yes,
plays games. He is currently working on a few demos and a few games,
as well as the "in-progress-but-sometimes-stalled-for-a-real-long-time-
until-inspiration-hits-again Internet stuff". He is also working on
raising a family, and enjoys music (particularly playing bass and guitar),
church, ice hockey and cricket, and classic video games.

Levente Harsfalvi is a 26 year old microcontroller programmer who works at
a small local company. His first C= encounter was a Plus/4 at school, at
the age of 12, and later (1988) his parents bought a C-16. After learning
BASIC and ASM coding he joined a Plus/4 demo group (GOTU, and later Coroners),
and has worked on game conversions, an FLI editor, music software (including
a SID emulator to play c64 music on TED) and numerous other software and
hardware projects. More recently he has begun taking some measurements on
the Plus/4 to figure out things such as how the sound generator works and is
working on a C-16 demo. Outside of the C= he enjoys cycling and running,
and ~50km walking tours.

For information on the mailing list, ftp and web sites, send some email
to chacking-info@jbrain.com.

While http://www.ffd2.com/fridge/chacking is the main C=Hacking homepage,
C=Hacking is available many other places including

        http://www.funet.fi/pub/cbm/magazines/c=hacking/
        http://metalab.unc.edu/pub/micro/commodore/magazines/c=hacking/

.... Jiffies ....

$01 Not _too_ long ago, an effort was made to write down the pin assignments
    for the video port of all the major C= computers. The result, as compiled
    by William Levak <wlevak@cyberspace.org>, is:

              8   7
Commodore   6   Video Connector
              3   1

            5   4
                 2
                                              Plus4   C16/C128
      CBM-II   VIC20   VIC20CR   C64   Pin C64A/SX64   C64B/C/E
   --------- ------------ ------- ---------- --- ---------- ----------
(R) Luminance +5 V   +5 V   Monochrome   1 Monochrome Monochrome (Y)
    Ground   Ground   Ground   Ground   2 Ground   Ground
(B) V. Sync. Audio   Audio   Audio   3 Audio   Audio   (W)
(W) Video   50 Ohm Video Video   Video   4 Video   Video
(Y) H. Sync. Video   Video   Audio In   5 Audio In   Audio In
                                                  6 Chroma   Chroma   (R)
                                                  7
                                                  8   +5 V

$02 "Professor Dredd" <profdredd@yahoo.com> has uploaded the programs
    from "Inside Commodore DOS" to his webpage at

        http://www.geocities.com/profdredd

$03 Todd Elliot has written a version of Pasi Ojala's zip/unzip program
    for GEOS/Wheels, available at:

        http://www.cs.tut.fi/~albert/Dev/gunzip-geos/

$04 Jeri has been hard at work on her video/cpu-board:

        http://www.geocities.com/cm easy/

$05 Jolse has been hard at work on JOS (but you'll have to wait until next
    issue for the article!):

    Here's the latest Jos news:
    Support for CMD HD, including native partitions. (Read only atm, and no
    1581 partitions)
    Enhanced shell with filename completion and recursive wildcards.
    Improvements to the GUI - the architecture now allows for a different
    window interface styles transparent to the application.
    Numerous showstopper bugs killed.
    Improvements to the httpd server to allow directory listings.
    Swiftlink/T232/Duart drivers.
    Started writing tutorials for programmers wanting to give Jos a go.

    Plus heaps more things not worth mentioning..

    cya!,
    Jolse

$06 Philip 'Pepto' Timmermann has made some very accurate measurements
    (and RGB calculations) of VIC-II colors:

        http://www.pepto.de/projects/colorvic/

$07 And finally, I have written a 2D graphics library for use by assembly
    language programs (plot points, draw lines and circles, that kind of
    thing). It's super-easy to use and pretty fast, so go ahead and use it
    if you need some hires drawing routines!

    For more information, pop on over to

        http://www.ffd2.com/fridge/grlib/

.... Side Hacking ....
                        Super/Turbo CPU VDC Hack
                                  for
        Commodore 128 with SuperCPU 64 or Turbo Master CPU Accelerators

                             by Henry Sopko
                         (henry.sopko@hwcn.org)

As many of you probably know, accessing the VDC (8563) chip is not
possible when using a SuperCPU 64 (version 1 tested only!) or the
TurboMaster CPU (latest revision), until now. With this 1 wire hack,
you can have access to the C128 (flat only tested) VDC 80 column
chip with your SCPU 64 or TurboMaster CPU!

DISCLAIMER

I take no responsibility whatsoever to any damage that may occur to
your Computer or SCPU 64/TM CPU resulting from this hardware hack!
You do this hack totally at YOUR OWN RISK!

[C=Hacking disclaimer: if you screw it up, it's your fault!]

This hack was only tested on a flat C128, using CMD's SCPU 64 (version 1).
Tested also was the Turbo Master CPU 4.09 MHz accelerator from Schnedler
Systems (latest revision).

Parts required: (1) long wire (12 inches or so, cutting it to length).

INSTRUCTIONS

Dissassemble your C128, taking the metal shield completely off. With the
motherboard exposed, find the 8563 VDC chip (U22). After locating this chip,
remove it (take note of the notch position, so you correctly re-insert the
chip). Bend out PIN 9 (R/W) of the 8563 just enough so it does not touch the
socket or any metal (in the flat 128, theres not much room, so be carefull).

Now re-insert the 8563 chip. Take a wire (you can either solder the wire to
pin 9 as I did, or use Microclips -- your choice). Now, connect the wire by

soldering or using Microclips to PIN 5 on the CARTRIDGE EXPANSION PORT. Thats
it!

I have done this hack quite a while ago without any signs of problems
whatsoever. The C128 functions the same with or without a SCPU 64 or TM CPU
connected after preforming this hack. I really did this hack for the Turbo
Master CPU so I could access the VDC. Turns out, that the SCPU 64 accelerator
(and maybe others?) work with this hack as well. The better choice of course
is to buy a CMD SuperCPU 128 to take full advantage of the Commodore 128 in
both modes!

::

                        16K Memory Expansion for the Vic-20

By Per Olofsson <MagerValp@cling.gu.se>.

Thanks to Ruud Baltissen, Pasi Ojala and Spiro Trikaliotis for help and
suggestions. The latest version of this project may be found at
http://www.cling.gu.se/~cl3polof/64.html.

I tried to keep the expansion as simple as possible, requiring no chips
besides the two memory chips, and a minimum of soldering. It uses two 6264
SRAM chips (62256 also works) piggy-backed to the Kernal and Basic 2364
ROMs. It's recommended that you do the expansion one chip at a time, as this
greatly simplifies troubleshooting. After the first chip checks out OK, do the
second one. I'll start out by explaining a few basic principles.

o Static Electricity

We'll be using CMOS RAM chips, and they are very sensitive to static
electricity. If you don't have an anti-static wrist strap, make sure you touch
a grounded surface before you touch the chips. Do not work sitting on shag
carpet in a mohair sweater while petting the cat :)

o Be Careful With ICs

ICs are sensitive to heat, and keeping the soldering iron for too long on a
pin can toast the chip. If you need to redo a solder joint on a pin, wait for
it to cool down first.

o Up, Down, and Pin Numbering

The top of a DIP IC (the kind used in Vic-20s) is marked with a small notch,
looking something like this:

```
         ___ ___
  1  = |     V     | =  8
  2  = |           | =  7
  3  = |           | =  6
  4  = |           | =  5
        `_____'
```
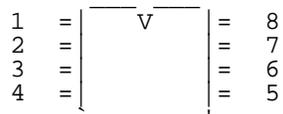
As you can see, the pin numbers start at 1, and the first pin is the top left,
going down to the bottom left, then from the bottom right to the top right.

Piggy-backing

Piggy-backing is a quick way of adding another IC on top of an existing
one. Normally you would create a PCB with chip sockets, connect it to the
expansion port and populate it with memory chips, but with piggy-backing you
simply add chips on top of internal ones. This works when the new chip's pins
has signals that are identical to, or closely matches, the layout of the one
it's being mounted on top of.

```
                                            6264 RAM

           2364 ROM                NC    1  = |   V   | =  28  Vcc
                                   A12   2  = |       | =  27  /WE
          ___ ___                  A7    3  = |       | =  26  CS2
  A7   1  = |   V   | =  24  Vcc    A6    4  = |       | =  25  A8
  A6   2  = |       | =  23  A8     A5    5  = |       | =  24  A9
  A5   3  = |       | =  22  A9     A4    6  = |       | =  23  A11
  A4   4  = |       | =  21  A12    A3    7  = |       | =  22  /OE
  A3   5  = |       | =  20  /CS    A2    8  = |       | =  21  A10
  A2   6  = |       | =  19  A10    A1    9  = |       | =  20  /CS1
  A1   7  = |       | =  18  A11    A0   10  = |       | =  19  D7
  A0   8  = |       | =  17  D7     D0   11  = |       | =  18  D6
  D0   9  = |       | =  16  D6
```
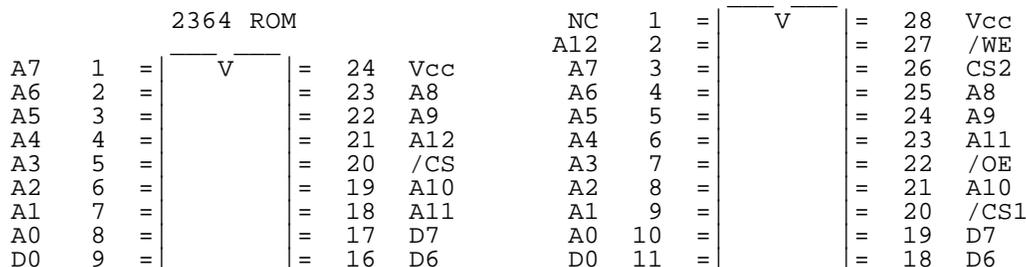
```
   D1  10  =|         |=  15  D5          D1  12  =|         |=  17  D5
   D2  11  =|         |=  14  D4          D2  13  =|         |=  16  D4
   Vss 12  =|_____  |=  13  D3          Vss 14  =|_____  |=  15  D3
           '-------'                               '-------'
```

As you can see there are 8 that don't match. However, we don't have to rewire
all of them. NC means "No Connection" so we can just ignore that pin. Note
that if you're using 62256 chips you'll have to wire this one too, see
below. We want to connect CS2 to Vcc, and we'll also swap A11 and A12, leaving
5 pins to solder on each chip. Swapping address bus pins works on RAM chips,
as it only affects how bits are stored internally -- when you read them out
again, they're swapped back. This also works for the databus.

We'll mount one 6264 on top of the Kernal ROM, and one 6264 on top of the
Basic ROM. The ROMs are marked UE11 and UE12 and can be found in the bottom
right of the motherboard. The wiring is identical for the two chips, except
for the /OE and /CS1 pins. The pins that we want to rewire we carefully bend
up so that they don't connect to the ROM. You want to bend them almost all the
way up (about 150 degrees) so that you can reach them with the soldering
iron. The pins are very sensitive, so make sure you bend the right pins --
bending a pin back again could easily break it.

Sometimes the pins on the RAM are too wide apart to make a good connection
when you piggy-back it. In this case, bend all the pins on the RAM slightly
inwards. You can do this by putting it on the side on a flat, hard surface and
press gently.

o Soldering A11/A12 and Vcc

You could get A11, A12, and Vcc from several places on the motherboard, but as
they're available on the ROMs we'll just solder small wires from the ROM to
the RAM. Remember that we're swapping A11 and A12, so connect pin 18 on the
ROM to pin 2 on the RAM. Connect pins 26 and 28 on the RAM to eachother.

o Soldering /WE

Pin 27 on the RAMs should be connected to pin 34 on the 6502 CPU. The CPU is
the 40-pin chip in socket UE10 on the motherboard, right next to the ROM
chips. Pin 34 is the 7th pin if you count from the top right pin on the CPU.

o /OE and /CS1

In the Vic-20 memory is divided into 8 blocks of 8K each. Block 0 is further
divided into 8 1K blocks, of which 5 are populated with RAM. Block 4 is used
by the VIC and the VIA chips, block 6 is the Basic ROM, and block 7 is the
kernal ROM. This leaves four blocks (1, 2, 3 and 5) available for cartridges
and RAM expansions. For RAM to be visible to the basic interpreter, you must
start by adding ram in block 1, then 2 and then 3. RAM in block 5 is never
detected by the basic. 8K cartridges use block 5, and 16K cartridges use block
5 together with another one, usually 3. To be as compatible as possible with
existing cartridges and to expand basic memory I'll use block 1 and 2 for this
expansion, but you could use any two of the available blocks you want. I've
added instructions for making the blocks selectable by switches below.

The block signals are available on the 74LS138 decoder chip marked UC5 on the
left side of the motherboard. Block 1 is on pin 14, block 2 is on pin 13,
block 3 is on pin 12 and block 5 is on pin 10. If you look closely you'll see
that the signals for block 1 and 2 go out from the chip a few mm to a small
pad. It's much easier to connect your wires to these pads than the pins on the
decoder chip. Solder a wire from block 1 to pin 20 and 22 on the first RAM
chip, and a wire from block 2 to pin 20 and 22 on the second RAM chip.

You're done!

That's all. When you power up the Vic-20, you should be greeted with a 19967
BYTES FREE message. If you've only done one chip so far, you'll get 11775
BYTES FREE, provided you connected it to block 1. If you've connected memory
to other blocks but not block 1, you'll just get the normal 3583 BYTES FREE
message. To test your memory, try this program:

```
  10 input "test which block";b
  20 s = b*8192 : t = s+8191 : e = 0
  30 print "testing databus"
  40 for a = s to t : poke a, 85 : if peek(a)  85 then gosub 1000
  50 poke a, 170 : if peek(a)  170 then gosub 1000
  60 next : de = e
  70 e=0 : print "testing high address bus"
  80 for a = s to t : poke a, int(a/256) : next
  90 for a = s to t : if peek(a)  int(a/256) then gosub 1000
 100 next : he = e
 110 e=0 : print "testing low address bus"
```

```
120 for a = s to t : poke a, a and 255 : next
130 for a = s to t : if peek(a)  a and 255 then gosub 1000
140 next : print
150 print de;"errors found in databus test"
160 print he;"errors found in high address bus test"
170 print e;"errors found in low address bus test"
999 end
1000 e = e+1 : print "error at";a : return
```

The program takes a couple of minutes to run.

Troubleshooting

  The computer doesn't start at all, or all you get is a black screen

   You've probably shorted or toasted something. Not good, this could
   have damaged the computer. Recheck all your soldering and make sure
   that you haven't accidentally connected something wrong.

  The computer powers up with 3583 bytes free

   Memory in block 1 is either not connected or not working. Check the
   connections between block 1, /CS1 and /OE, between R/W and /WE, and
   between Vcc, CS2 and Vcc.

  The computer powers up with something other than 3583, 11775 or 19967 bytes
  free

   Memory is functioning partially, check A11 and A12. This could also
   indicate that a RAM chip is faulty.

  The computer powers up with the correct number of bytes free, but the memory
  test program fails

   Memory is functioning partially. If the databus test fails, check the
   connections between block 1, /CS1 and /OE, between R/W and /WE,
   between Vcc, CS2 and Vcc, and D0 through D7. If the address bus test
   fails, check A0 through A12.

Using 62256 chips instead of 6264

You can freely substitute 62256 chips for the 6264. Only two pins differ:

```
              6264 RAM                              62256 RAM

    NC    1   = |‾‾‾‾‾| =  28  Vcc       A14   1   = |‾‾‾‾‾| =  28  Vcc
   A12    2   = |  V  | =  27  /WE       A12   2   = |  V  | =  27  /WE
    A7    3   = |     | =  26  CS2        A7   3   = |     | =  26  A13
    A6    4   = |     | =  25  A8         A6   4   = |     | =  25  A8
    A5    5   = |     | =  24  A9         A5   5   = |     | =  24  A9
    A4    6   = |     | =  23  A11        A4   6   = |     | =  23  A11
    A3    7   = |     | =  22  /OE        A3   7   = |     | =  22  /OE
    A2    8   = |     | =  21  A10        A2   8   = |     | =  21  A10
    A1    9   = |     | =  20  /CS1       A1   9   = |     | =  20  /CS1
    A0   10   = |     | =  19  D7         A0  10   = |     | =  19  D7
    D0   11   = |     | =  18  D6         D0  11   = |     | =  18  D6
    D1   12   = |     | =  17  D5         D1  12   = |     | =  17  D5
    D2   13   = |     | =  16  D4         D2  13   = |     | =  16  D4
   Vss   14   = |_____| =  15  D3        Vss  14   = |_____| =  15  D3
```

Pin 26 (A13) can be connected to Vcc just like on the 6264, but we need to
wire pin 1 (A14) to either Vcc, Vss or another address bus pin. It's probably
easiest to wire it to pin 2 (A12). We won't be using the greater capacity of
the 62256 using this method (doing so would require some kind of decoder
logic) but sometimes 62256 chips are cheaper than 6264, or maybe you just have
some lying around.

Adding Block Select and Write Protect Switches

Adding block select switches allows you to chose which blocks should be
populated on the fly. This allows you to chose between a basic expansion and a
"cartridge emulator" that allows you to load cartridge images into ram and run
them. I added one switch for each chip, giving the first chip the option
between block 1 and 3, and the second between 2 and 5.

```
  block 1 -----o
                \
```

```
                 o----- to /CS1 and /WE on chip 1

  block 3 -----o


  block 2 -----o
                \
                 o----- to /CS1 and /WE on chip 2

  block 5 -----o
```
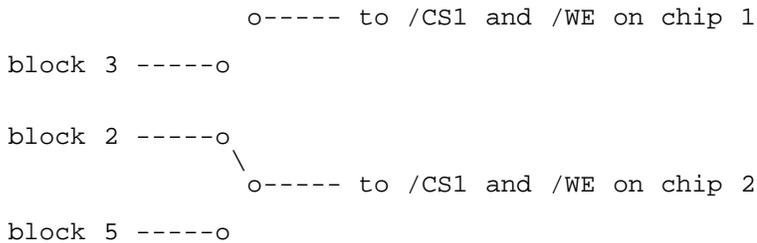
As a kind of copy protection, some cartridge images try to modify themselves
to detect if they're running from RAM. If we add write protect switches we'll
be able to run those as well. Switch to write enabled while loading, then
switch to write protected and reset.

```
       R/W -----o
                \
                 o----- to /WE on chip 1

      +5 V -----o


       R/W -----o
                \
                 o----- to /WE on chip 2

      +5 V -----o
```
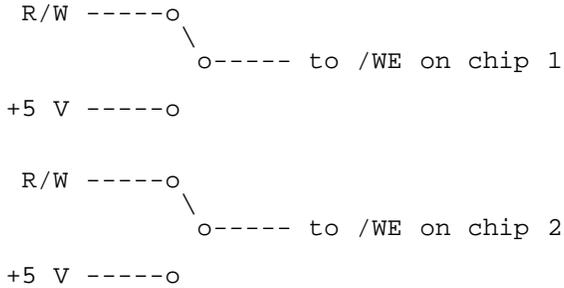
+5 V is the same as Vcc.

:::

Quick Quadratic Splines
--------------------- S. Judd <sjudd@ffd2.com>

Splines are neat.

Splines are a way of drawing curves -- specifically, drawing a curve through
any set of _points_ that you choose; for example, a curve that starts at (0,0),
goes over to (100,50), then loops back to (50,0), and continues on through
any number of points can be drawn just by specifying the points.

This is a very useful thing! One place splines are used is in drawing fonts.
With a relatively small list of points, it becomes possible to draw letters
in arbitrary shapes (and scale those shapes easily). Another application
is animation -- it is possible to specify a long motion path using just
a few points along the path. Furthermore, an animated object might have
different parts that move differently -- arms, legs, head, etc. By just
specifying a few frames of the animation, splines can be used to generate
the in-between frames.

Chances are good that by now you are thinking about some of your own
applications for splines, so let's see how they work.

Most articles/books/etc. that talk about splines talk about "cubic splines".
This article is going to talk about "quadratic splines", which never seem
to be mentioned. Which is a pity, since quadratic splines are perfectly
adequate for many (if not most) spline applications and are far more
computationally efficient.

Graphically, a two-dimensional quadratic spline looks something like:

```
             * P1
            / \
           /   \
          /     \
         /  . .  \
        / .     . \
       /.         .\
      /.           .
     /.              * P2
    .
     .
P0 *
```
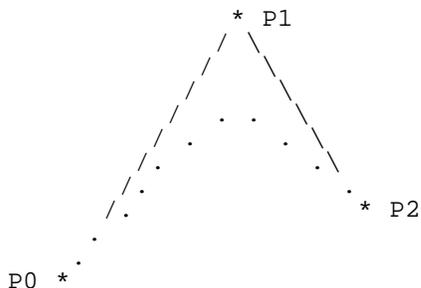
(Another brilliant display of ASCII art). The curve starts at P0, moves
towards P1, and then turns around and heads towards P2, where it ends. If
we draw lines from P0 to P1 and from P1 to P2, these lines are tangent to
the curve at the endpoints P0 and P2; that is, these lines say what

"direction" the curve is going at the endpoints. Using the above diagram,
it should be very easy to visualize how changing P1 changes the shape of
the curve.

So here, for your computing pleasure, is a quadratic spline:

P(t) = P0*(1-t)^2 + 2*P1*t*(1-t) + P2*t^2

where P0, P1, P2 are constant values, and t ranges from zero to one. If
you don't like equations, then how about a computer program to compute
the above:

```
10 P0=10:P1=100:P2=37
20 FOR T=0 TO 1 STEP 1/16
30 PRINT T,P0*(1-T)*(1-T) + 2*P1*T*(1-T) + P2*T*T
40 NEXT
```

This is actually very easy to understand. When t=0, P(t) = P0. As t starts
to increase, (1-t) starts to decrease and P(t) starts moving towards P1. As
t gets even larger P(t) starts moving towards P2, until finally, at t=1,
P(t) = P2.

You may be wondering why there is that 2*P1 in the equation, instead of
just P1, but that will become apparent shortly.

At this point there is an important thing to notice about the above
equations: adding _any_ constant to P0 P1 and P2 is just like adding the
constant to the entire spline P. Let P0=P0+C, P1=P1+C, P2=P2+C; the
spline is then

P'(t) = (P0+C)*(1-t)^2 + 2*(P1+C)*t*(1-t) + (P2+C)*t^2
      = [P0*(1-t)^2 + 2*P2*t*(1-t) + P2*t^2] + C*[(1-t)^2 + 2*t*(1-t) + t^2]
      = P(t) + C

This is one reason that factor of 2 multiplying P1 is important -- it makes
the expression multiplying C add up to 1. This property means that splines
can be _translated_ easily. Note that one value you can translate everything
by is P0 -- subtract P0 from each point and the spline becomes

P(t) - P0 = 2*(P1-P0)*t*(1-t) + (P2-P1)*t^2

and hence

P(t) = 2*(P1-P0)*t*(1-t) + (P2-P1)*t^2 + P0

which gets rid of the P0*(1-t)^2 term. Depending on how the spline algorithm
works, this can save some computation time.

Now, the points P0, P1 etc. are called "control points". If you're on the ball
(or if you've tried the BASIC program above) you've noticed that while the
spline starts at P0 and ends at P2, it never actually reaches P1 -- it just
heads towards it but then heads away towards P2. If a different P1 is
chosen, the spline still starts at P0 and ends at P2 but takes a different
path between the endpoints. P1 controls the shape of the curve.
So now let's say we want a curve that passes through three points, s0, s1,
and s2, using a quadratic spline. P0 and P2 are easy to choose:

        P0 = s0
        P2 = s2

P1 can actually be chosen at this point to be any number of points. All we
do is choose a value for t -- call it t1 -- when the spline should hit s1:

P(t1) = s1 => P1 = (s1 - P0*(1-t1)^2 - P2*t1^2) / (2*(t1*(1-t1))

(just substitute t=t1 into the equation for P(t) and solve for P1). So if
t1=1/2, say, then

P1 = (s1 - P0/4 - P2/4) / (2 * 1/4)
   = 2*s1 - (P0 + P2)/2

and the spline will hit s1 halfway through the iteration at t=1/2.

In a typical application, of course, the curve will need to go through tens,
hundreds, even thousands of points. Naturally, the way to do this is by
joining a bunch of splines together.

If we just choose a spline for every three points, using the above equation
for P1, there will typically be sharp corners where the splines join
together. Sometimes this is a good thing -- for example, when drawing
something like a valentine we might want sharp corners. But usually

what is needed is a nice smooth curve through all the points.

With quadratic splines this is a piece of cake. Each spline has three
control points: the two endpoints, and the middle control point P1. The
endpoints are simply chosen to be the points we want to pass through,
and P1 can then be chosen to make all the connections smooth.

That is, given a list of points s0, s1, s2, ... to put a curve through,
choose s0 and s1 to be the endpoints of the first spline, s1 and s2 to
be the endpoints of the second spline, and so on, and choose the middle
control points to make everything smooth.

"Smooth" here means a continuous first derivative -- if you don't know what
a derivative is then don't worry about it, it just means that the "slope" of
each spline is the same where they join together. Given a spline

S1(t) = P0*(1-t)^2 + 2*P1*t*(1-t) + P2*t^2

the derivative at t=1 is given by

S1' = 2*(P2 - P1)

P2-P1 is simply the line segment running from P1 to P2, which verifies what
was said earlier: at P2 the curve is tangent to a line drawn from P1 to P2.

```
                    * P1
                  / \
                 /   \
                /     \
               /       \
              /  . .   . \
             / .       . \
            /.           .\
           /.             .
          /.               . * P2
         .
        .
      P0 *
```
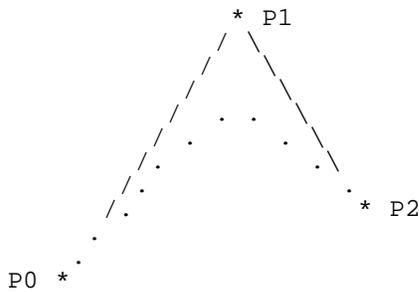
P2-P1 is the direction the curve is headed in. So to join another spline
smoothly to the first one, simply extend this line segment and make sure
the second middle control point lies somewhere on that line:

```
                  * P1
                / \
               /   \
              /     \
             /       \
            /  . .   . \
           / .       . \
          /.           .\
         /.             .
         /.              . * P2
        .                 .         * P5
       .                   \.     ./
     P0 *                    \  ../
                              \ /
                               * P4
```
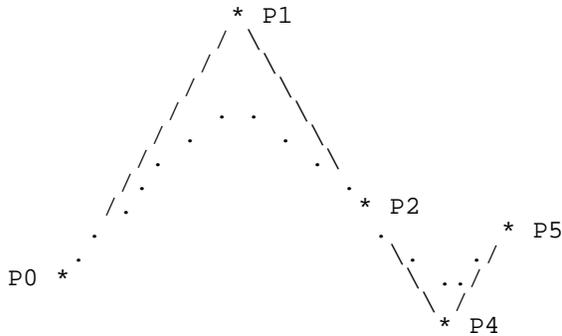
Since the first spline is tangent to P2-P1, and the second spline is tangent
to P4-P2, the two splines have the same slope at P2, and join smoothly.
Note that P4 can be _anywhere_ along the P2-P1 line, and choosing different
values for P4 will change the curve.

To make this more precise, let the two splines be given by

S1(t) = P0*(1-t)^2 + 2*P1*t*(1-t) + P2*t^2
S2(t) = P3*(1-t)^2 + 2*P4*t*(1-t) + P5*t^2

To join them smoothly, the slope of the first spline must be proportional
to the slope of the second spline at the joint:

P4-P3 = c*(P2-P1)

where c is some constant (the _direction_ of the two slopes need to be the
same, but not the magnitude). So, to make the second spline fit smoothly
to the first we simply choose the middle control point P4 as

P4 = P3 + c*(P2-P1)

or, in plain English, starting from P3 (the joining point) move some
distance c in the direction P2-P1. To see why the factor c is important
just scroll up a page or two to the diagram, and imagine how the curve

changes as P4 slides back and forth along the line.

We now have the tools to construct a smooth curve that passes through any
points s0, s1, s2, s3, ... Let the first spline pass through s0, s1,
and s2, using the t=1/2 formula given earlier:

```
P0 = s0
P1 = 2*s1 - (s0 + s2)/2
P2 = s2
```

P1 could be chosen in other ways, of course, but the above usually works
pretty well. Then choose each succeeding spline to match up smoothly with
the previous spline:

```
P3 = s2                    ;start at spline 1 endpoint
P4 = P3 + c1*(P2-P1)       ;smooth joint
P5 = s3                    ;next point

P6 = s3
P7 = P6 + c2*(P5-P4)
P8 = s4
```

and so on. (When drawing smooth curves, it is of course really only
necessary to store the middle control points P4 P7 etc. for each spline).
The choice of the constants c1 c2 etc. really depends on how you want the
curve to look.

As you can see above, each extra point requires another spline. One spline
per point may sound like a lot, but there are a lot of pixels in-between each
pair of points, and more importantly these are quadratic splines and hence
can be made to go very, very fast.

Computing the spline

Note that the control points are one-time calculations; those three constant
values completely determine the spline. One way of drawing the spline is
to use some lookup tables for t^2, $2*t*(t-1)$, and $(1-t)^2$ and do a little
fast multiply magic. This is fast and straightforward, so I won't talk
about it much except to mention that signed multiplies may be required,
and each curve is limited to 256 points (t can go from 0 to 255 in the
lookup tables). Note also that it would be nice to have an "incremental"
plot routine, i.e. one that just updates pointers when necessary, instead
of recomputing all the bitmap pointers at every iteration.

Which of course brings us to a second way of drawing the spline: to ask
"how does the spline change when I increment t?"

Let's say that t is incremented by dt at each step (t -> t+dt). At each
step, then, the spline changes by

$S(t+dt) = P0*(1-t-dt)^2 + P1*2*(t+dt)*(1-t-dt) + P2*(t+dt)^2$

$= S(t) + k1*dt^2 + k1*2*t*dt + k2$

where k1 = P0 - 2*P1 + P2 and k2 = 2*dt*(P1 - P0). That is, to advance one
step, we add

$k1*dt^2 + 2*k1*t*dt + k2$

to the current value. Take a look at how the above value changes as t
changes:

```
t       k1*dt^2 + 2*k1*t*dt + k2
--      -----------------------
0       k1*dt^2              + k2 =   k1*dt^2 + k2
dt      k1*dt^2 + 2*k1*dt^2 + k2 = 3*k1*dt^2 + k2
2*dt    k1*dt^2 + 4*k1*dt^2 + k2 = 5*k1*dt^2 + k2
3*dt    k1*dt^2 + 6*k1*dt^2 + k2 = 7*k1*dt^2 + k2
```

and so on. This means that all we have to do to advance the spline is
something like

```
        c0 = k1*dt^2
        C = c0 + k2
        S = P0
:loop
        S = S + C
        plot S
        C = C + 2*c0
        loop
```

Which is really, really fast. Instead of specifying the spline with the
three values P0, P1, and P2 we can specify it by the three values P0, k1,
and k2.

It is possible to understand this iterative method somewhat. The update
constant k1 can be written as

        k1 = (P0-P1) + (P2-P1)

whereas k2 goes like

        k2 ~= (P1-P0).

P1-P0 is the "direction" line segment towards P1, so the curve starts moving
towards P1. But k1 starts to balance that out with the P0-P1 term (the line
segment pointing in the opposite direction), while moving towards P2 with the
P2-P1 term, which is the direction line segment from P1 to P2. Now, isn't
that all nice and clear now?

The Program

At the end of this article I've included a sort of "spline laboratory" --
it's a little BASIC program that lets you experiment with different aspects
of splines and see what they look like. It uses BLARG for the graphics, so
I've included that as well.

When you run it, it plots three points on the screen to draw a curve through.
You can move these points around, as well as add new points, draw the splines,
draw the control point "a-frames", and so on. It has a ton of commands,
basically because I just added them as I got curious about different things:

+/-   Move to next/previous point
cursor keys   Move current point
return   Draw spline
space   Clear screen
*   Toggle point display
< >   Decrease/increase accuracy (number of bits in iteration)
1 <-   Increase/decrease time step (number of points in spline)
=   List control points (numerical values)
.   Draw control point a-frame
S   Toggle smooth splines (see below)
A   Add more points
Q   Quit

Pressing "A" adds three points at a time, starting at the last point in
the curve. The reason for adding three points, instead of just one, is
because of the "S"mooth spline feature. When smooth splines are selected,
the program will draw a smooth curve through all the points. When toggled
off, the program will draw a spline through every three points, using the
"t=1/2 method" to select the middle control point.

The program uses the fast iteration method described earlier, using fixed
precision arithmetic (since this is what an assembly program would do).
The < and > keys are used to change this precision, since certain splines
need more precision than others.

Finally, the value CC=0.4 is set in line 10. This is the "smoothing"
constant, i.e. c1 in the equation for P4 below:

P3 = s2   ;start at spline 1 endpoint
P4 = P3 + c1*(P2-P1)   ;smooth joint
P5 = s3   ;next point

You might want to experiment with different values, to see what happens
(or else let each spline have their own value).

Anyways, the program is not meant to be terribly profound -- just a starting
point and something to play with to work out your own ideas on!

Cubic Splines

Just for completeness, here is a cubic spline:

P(t) = P0*(1-t)^3 + 3*P1*t*(1-t)^2 + 3*P2*(1-t)*t^2 + P3*t^3

As you can see, it has four terms instead of three, and is cubic in t. But
the idea is the same -- at t=0 it starts at P0, as t increases moves
towards P1, then towards P2, and finally ends at P3.

You can also see that it is significantly more computationally involved than
the quadratic spline. Moreover, computing the middle control points P1 and P2
is also fairly involved -- much more involved than with the quadratic spline.

So, why use a cubic spline? Two reasons: first, you can specify the derivative
at _both_ endpoints -- the direction line P1-P0 will be the slope at the
starting point, and the direction P3-P2 will be the slope at the end point.
This gives some flexibility needed for certain applications.

Second, you can alternatively match _second_ derivatives at the connection
points, which again can be useful for certain applications.

For more information about cubic splines, just check out any decent graphics
book.

The Program

```
begin 644 slab.zip
M4$$L#! H     & $]LQQJSS3>D,0>   /O(     *        8FQA<F<N;6TN;P\    0(4
M!18%!!@@@@!@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
```
(uuencoded binary data continues — not reproducible as readable text)

```
MAE8)K"Z3M.+K(%P*:<O1,M*NL2&8<);):EQ7"32O8<LI^ZK0<VRI)Q\9?Y9Q
M\"SCG&6< ?)LJ 9V"6@R!T3FO.%;@HLNWF$PYC%KY'JS#>G,$VE@J'2Y#?T2
M '487QVJ'6_#OP1U K3<$# !F0;VFJ7VJD^K3E%^5:E6BS"L-PPTWMH-!Q.
MY94L6=1CRW)B&DZ_WA"1Y@!.HCED^0 [.FQ,0%&6Y2GF4&RT)>NJ8E\MHV F
M<,T"#L\"KEG <8!%'4H"/ Y3/*[*: -\U>$EA=+E)8$,RRHIUR DV5_):T3I\
MM5@#.X3U6HD694HU:,NF16;[4R*ML%GLUSN,Y5W)#FL3H 6W=JB; "U89],@
MI+B-B_N9-EN1M+92W"&QR@X#,@SPQY05&BE?IJQ4T,PTALY0LG"?B/^Q7)](
M(9[QH;-ZS55**/9JD'SU(/FJ5>L]3KTTH?80E;4\"S\3 5Z4BN>\\2OD\CR.F\K
MY%&9!9Q0@&\"?2B>0,B$RRJ#]](%.QU 6H]B$[@84#!O',OO0,O!]A(>RCL5F^,."!6+F
M+3_6OE4ANVKKJM++"\W'T-EXGCX"W^HP%QGGPQA'58WSJG$\Q\0\YP\I%A@\
MPKKL:V>T_] #O6:Y@V'8EM..<*G&_&(::H2(PERN,,'X9IRL,I7.8$$%L6@@"W"LUF/
M,9L);V,M9#I/B@#(+N#2+#/#*6<!B0$ZA25D>7D5^M$"T6N/5K(3X_%.lll]I-H
M[ =TDMG#!'Z 8I^!.5;U=3CU=3C)\"H:"L??R&1&&1WRGG5/4[$/4?L"CD[53926\
MA'.(Y ^(=)(R!EIW0W1_H*136<X%!.!PB_O.H"0+[;;;>]>[;L-.,#3AGYV?2^'Z+P
MA41>XBGE-H\"CU#;;;)KA-(9NG4=D3/DD\"P"*$\$9E]KWK8NNB47LLLC0E^G?3Q&%&<.
M^S,,S,,C6L->>5_3(T)".J",VW[.03'T4SI#@*<&A'<O==& 1Y+&&)C3W-E-., LT94
MO])#1AA5O1P$"""""]@.* !@!/;,<:L\TWI#$'  #T"  "@
MI($     8FQQ<F<<N;6TN;U!+   0(* PH   S &&1LQQ%;@@;  %)P 
M                "D@5L'   !S<&QI;FB.8L32+04Lll&      (  @!O     C@T
#
```

end

....
....
..
.   C=H #20

:::

Main Articles:

:::

VIC KERNAL Disassembly Project - Part IV
Richard Cini
February, 2001

Introduction
============

        In the last installment of this series, we examined the six remaining
routines that are called from the IRQ and NMI vectors. The routines examined
are responsible for updating the jiffy clock, determining the location of
color RAM, scanning the keyboard, and reading and setting the cursor position.
        Having fully completed the main processor vectors, we'll continue this
series by examining other Kernal routines.

More Subroutines
================
        Much of the last three articles dealt with routines that were used in
the VIC's startup process, configuring the screen and I/O devices, and polling
the keyboard for user input.

        Although the entire Kernal ROM consists of probably over 100
individual subroutines -- all indirectly callable in some way - the VIC's
developers really intended the user to only call the 39 routines accessible
through the jump table at the end of the ROM. When looking at this table,
we've so far only discussed only 10 of the routines contained therein.

        It was designed in this fashion for multiple reasons. The functions
selected for the jump table were clearly the ones Commodore found to be most
useful for software programmers. The other routines were "internal" routines
used by the "public" routines and the BASIC ROM.

The jump table also hid code changes from the software programmer, although I
don't believe that there were ever any significant updates to the Kernel after
its initial release. The calling routine and parameters remained consistent
even if the underlying code changed. It also facilitated sharing programs
between machines with common lineage. For example, certain programs could be
shared between the PET, VIC and the C64. Of course, minor modifications of
these programs were required to account for different I/O locations.

This structure is no different than a modern operating system such as Windows,
or even the Macintosh, which had the concept of a ROM "toolbox."

Anyway, I digress...

When examining the functions available in the jump table, the following
distribution becomes apparent:

```
User I/O (screen/keyboard):        3 functions
Timekeeping:                       3 functions
Memory management:                 3 functions
Processor control:                 4 functions
Device I/O:                       26 functions
```

Commodore appeared to be giving greatest support for interfacing the VIC to various devices. Many of these routines were used by the built-in BASIC interpreter to support file I/O through the tape deck and the serial IEEE port, as well as printer and RS232 support. Unfortunately, the only devices that made it onto the IEEE bus were floppy drive and printer devices. If there are others, I appreciate knowing about them.

        Anyway, this provides a nice segue into looking at the device I/O routines in some detail.

        The first step is opening a device. All devices are eligible to be opened in this manner. The OPEN function requires that the program call two preparatory routines first. The first one, SETNAM sets the filename parameter required by OPEN. Clearly only certain devices make use of a filename (such as the cassette or disk drive devices). For devices not requiring a filename pointer, just set the pointer to "null".

```
FE49   ;===================================================
FE49   ; ISETNM - Set filename (internal)
FE49   ; Call with .A = filename length, .X = LSB of filename
FE49   ; string, .Y = MSB of filename string
FE49
FE49   ISETNM
FE49 85 B7      STA FNMLEN              ; set length
FE4B 86 BB      STX FNPTR               ; ptr L
FE4D 84 BC      STY FNPTR+1             ; ptr H
FE4F 60         RTS
```

        Calling SETNAM sets three Zpage variables that tell the OPEN routine where in memory to find a string containing the filename. Setting the length parameter to zero yields a null string.

        When opening an RS232 channel, the filename string would contain two ASCII characters representing the configuration values for the command register and the control register.

        The next call that is required is to set the underlying device parameters for the OPEN call.

```
FE50   ;===================================================
FE50   ; ISETLF - Set logical file parameters (internal)
FE50   ; Call with .A = file number, .X = device # (0-30),
FE50   ; .Y = command
FE50   ISETLF
FE50 85 B8      STA LOGFIL              ; file #
FE52 86 BA      STX CHANNL              ; device
FE54 84 B9      STY SECADR              ; secondary address
FE56 60         RTS
```

        The SETLFS parameters are fairly common: the file handle number, the device to access, and any secondary address to send to the device. The combination of SETNAM and SETLFS is analogous to this BASIC statement:

        OPEN 2,8,1, "mydat"

where "2" corresponds to the file handle number, "8" is the device, and "1" is the secondary address. These numbers are ultimately stored in their respective data tables at location IOPEN_S3 in the OPEN routine.

        Opening the RS232 channel does not require the use of a secondary address.

        During file I/O, it's helpful to check the status of the current device by using the READST function. READST retrieves the value of the CSTAT variable. CSTAT is set through a call to ISETMS1 (at $FE6A) by various I/O routines based on device response.

```
FE57   ;===================================================
FE57   ; IRDST - Get I/O status word (internal)
FE57   ;
FE57   IRDST
FE57 A5 BA      LDA CHANNL              ; get current device
```

```
FE59 C9 02        CMP #$02                    ; RS232?
FE5B D0 0B        BNE ISETMS+2                ; no...branch to OS messages
FE5D AD 97 02     LDA RSSTAT                  ; get RS232 status
FE60 A9 00        LDA #$00
FE62 8D 97 02     STA RSSTAT
FE65 60           RTS
FE66
FE66
FE66    ;=================================================
FE66    ; ISETMS - Control OS messages (internal)
FE66    ; On entry, .A is message number. Bit7=1 for KERNEL messages, and
FE66    ; Bit6=1 for control messages. Bit0-Bit5 is message number.
FE66    ;
FE66    ; set flag for OS messages
FE66    ;
FE66    ISETMS
FE66 85 9D              STA CMDMOD              ;save message #
FE68 A5 90              LDA CSTAT               ;get status
FE6A
FE6A    ; set ST bits
FE6A    ISETMS1
FE6A 05 90              ORA CSTAT               ;twiddle bits based on .A
FE6C 85 90              STA CSTAT               ;save status
FE6E 60                 RTS
```

The return values of the READST call have different meanings depending
on what device is being accessed:

BIT	Cassette	Serial R/W	Tape L/V
0	Write timeout		
1	Read timeout		
2	Short block	Short block	
3	Long block	Long block	
4	Unrecoverable	Any mismatch read error	
5	Checksum error		Checksum error
6	End of file	EOI line	
7	End of tape	Device not present	End of tape

So, after having set the file name information and setting the device
access parameters, we're ready to actually OPEN the file. The following code
is heavily commented, so I'll only expand where necessary.

The Kernal keeps track of the open files using a series of file handle
tables. A maximum of 10 logical files can be open at one time. FILTBL keeps
track of the logical file handle established in the call to SETLFS. SECATB
keeps track of the secondary address associated with a given logical file
number. Finally, the DEVTBL variable tracks which device number relates to the
file handle. These three tables mirror the variables in the SETLFS call.

```
F40A    ;=========================================================
F40A    ; IOPEN - Open file (internal)
F40A    ; Required prior calls:  FFBA/SETLFS and FFBD/SETNAM
F40A    ; No arguments
F40A    IOPEN
F40A A6 B8              LDX LOGFIL       ;get file number
F40C D0 03              BNE IOPEN_S1     ;F411 read from output file?
F40E 4C 8D F7           JMP IOERMS6      ;Yes, emit "NOT INPUT FILE" error
F411
F411         IOPEN_S1
F411 20 CF F3           JSR FIND         ;locate file # in table
F414 D0 03              BNE IOPEN_S2     ;file number not found, so move on
F416 4C 81 F7           JMP IOERMS2      ;found, so emit "FILE OPEN" error
F419
F419         IOPEN_S2
F419 A6 98              LDX COPNFL       ;get # of open files
F41B E0 0A              CPX #$0A         ;are there 10 files open already?
F41D 90 03              BCC IOPEN_S3     ;no, OK to open new file
F41F 4C 7E F7           JMP IOERMS1      ;more than 10, "TOO MANY FILES" error
F422
F422         IOPEN_S3                    ; allocate file slot in table
F422 E6 98              INC COPNFL       ;add 1 to count of open files
F424 A5 B8              LDA LOGFIL       ;get file number from call
F426 9D 59 02           STA FILTBL,X     ;save file # in table
F429 A5 B9              LDA SECADR       ; get secondary address
F42B 09 60              ORA #%01100000   ;$60 make it a device command
F42D 85 B9              STA SECADR       ;save it as secondary and add it to the
F42F 9D 6D 02           STA SECATB,X     ; SA table for that open file
F432 A5 BA              LDA CHANNL       ;save device number to the
```

```
F434 9D 63 02           STA DEVTBL,X     ;  device table
F437
F437                                     ;Special handling for certain devices
F437 F0 5A              BEQ IOPENRC      ; keyboard device? Yes, return
F439
F439 C9 03              CMP #$03         ;Is the device the screen?
F43B F0 56              BEQ IOPENRC      ; yes, then return clear
F43D 90 05              BCC IOPEN_S4     ;must then be the tape or RS232-branch
F43F
F43F                                     ; Start IEEE stuff
F43F 20 95 F4           JSR SENDSA       ;send secondary to IEEE bus and
F442 90 4F              BCC IOPENRC      ; return clear
F444
F444        IOPEN_S4
F444 C9 02              CMP #$02         ;RS232 (2)?
F446 D0 03              BNE IOPEN_S5     ;not RS232, so process tape device
F448 4C C7 F4           JMP SEROPN       ;open RS232 device
F44B
F44B        IOPEN_S5                     ;Tape device (1)
F44B 20 4D F8           JSR GETBFA       ;Get tape buffer address
F44E B0 03              BCS IOPEN_S6     ;MSB>2? No, continue with open
F450 4C 96 F7           JMP IOERMS9      ;Bad tape buffer, emit "ILLEGAL DEVICE
F453                                     ; NUMBER" error
F453
F453        IOPEN_S6                     ;Continue with tape processing
F453 A5 B9              LDA SECADR       ;get SA
F455 29 0F              AND #%00001111   ;Are we in write/save mode?
F457 D0 1F              BNE IOPEN2       ;yes, prompt for Play + Record
F459
F459 20 94 F8           JSR PLAYMS       ;wait for "PLAY" key
F45C B0 36              BCS IOPENRC+1    ;$F494 return CY=1
F45E
F45E 20 47 F6           JSR SRCHMS       ;print "Searching for [name]" message
F461 A5 B7              LDA FNMLEN       ;get length of filename
F463 F0 0A              BEQ IOPEN1       ;no name specified, so search for
F465                                     ;  next header
F465
F465 20 67 F8           JSR LOCSPH       ;search for specific tape header
F468 90 18              BCC IOPEN3       ;$F482 go get header
F46A F0 28              BEQ IOPENRC+1    ;$F494 return found CY=1
F46C
F46C        IOPENA                       ; done searching and file was not found
F46C 4C 87 F7           JMP IOERMS4      ;F787 "FILE NOT FOUND" error
F46F
F46F        IOPEN1                       ; Get next tape header
F46F 20 AF F7           JSR LOCTPH       ;search for next header
F472 F0 20              BEQ IOPENRC+1    ;$F494 return found CY=1
F474 90 0C              BCC IOPEN3       ;$F482 go get header
F476 B0 F4              BCS IOPENA       ;$F46C not found
F478
F478        IOPEN2                       ; write tape header
F478 20 B7 F8           JSR RECDMS       ;wait for REC & PLAY keys
F47B B0 17              BCS IOPENRC+1    ;$F494 return CY=1
F47D
F47D A9 04              LDA #$04         ;control byte ID for data header
F47F 20 E7 F7           JSR WRTPHD       ;write tape header
F482
F482        IOPEN3
F482 A9 BF              LDA #$BF         ;pointer to tape buffer head
F484 A4 B9              LDY SECADR       ; get secondary address
F486 C0 60              CPY #$60         ;SA=0 (read) mode?
F488 F0 07              BEQ IOPENRC-2    ;$F491 Yes, skip write
F48A
F48A A0 00              LDY #$00         ; write mode
F48C A9 02              LDA #$02         ;control byte ID for data block
F48E 91 B2              STA (TAPE1),Y    ;write it to buffer
F490 98                 TYA
F491 85 A6              STA BUFPNT       ;save pointer
F493
F493        IOPENRC
F493 18                 CLC
F494 60                 RTS
```

        The OPEN function spends most of its time saving parameter information
into variable tables and figuring out which device to open. If the OPEN call
is for the tape device, the code either reads or writes the respective tape
header.

        OPEN calls nine other routines, six of which deal with tape input and
output. The three non-tape routines are FIND (find file number in table),
SENDSA (send secondary address), and SEROPN (open RS232 device). Since the

OPEN code spiders out so, I'm going to leave the tape-related routines to
another article.

        FIND is a very easy routine. It searches through the file handle table
to see if the handle of the file we're trying to open is already being used as
a result of a previous call to OPEN. Clearly each file handle number has to be
unique to prevent confusion at the system level. If a file is already in use
with the same handle number, .X returns non-zero, enabling the calling routine
to drop into an error handler ("FILE OPEN" error).

```
F3CF      ;===============================================================
F3CF      ; FIND - Look for logical file number in open-file table
F3CF      ; On entry to FIND1, .A=file#. On exit, .X=offset in file table
F3FC      ; to matching file number.
F3CF
F3CF                FIND
F3CF A9 00          LDA #$00
F3D1 85 90          STA CSTAT        ;clear Status variable
F3D3 8A             TXA              ; .X is logical file number copied to
F3D4                                 ; .A for use in FIND1
F3D4                FIND1
F3D4 A6 98          LDX COPNFL       ;get #of open files to count through
F3D6
F3D6                FINDLOOP         ; loop
F3D6 CA             DEX
F3D7 30 15          BMI FLATRBX      ;$F3EE reached 0, then exit
F3D9 DD 59 02       CMP FILTBL,X     ;is this the one?
F3DC D0 F8          BNE FINDLOOP     ;$F3D6 no, try again
F3DE 60             RTS              ;return with .X= offset into table
```

        The FLATRB routine is not used in OPEN, but the FIND subroutine exits
through it. I don't know why this is since the FIND routine has a return path
of its own. Anyway, FLATRB takes an index number into the open file table and
sets the Zpage file variables according to the values pointed to by the
index. For example, if I call FLATRB with .X=0, the file variables will be set
with the information stored in location [0] of each of the file handle table,
the device table, and the secondary address table. So, one could select an
open file handle, use FIND to get the index number, and then use FLATRB to set
the Zpage variables to the right values for subsequent use.

```
F3DF      ;===============================================================
F3DF      ; FLATRB - Set file values
F3DF      ; On entry, .X = offset in the file tables. Returns with
F3DF      ; Zpage file variables set.
F3DF
F3DF                FLATRB
F3DF BD 59 02       LDA FILTBL,X
F3E2 85 B8          STA LOGFIL               ;get file handle number
F3E4 BD 63 02       LDA DEVTBL,X
F3E7 85 BA          STA CHANNL               ;get device
F3E9 BD 6D 02       LDA SECATB,X
F3EC 85 B9          STA SECADR               ;get SA
F3EE
F3EE                FLATRBX
F3EE 60             RTS
```

        After the OPEN routine determines that the file handle to be opened is
unique, assigns the file variables to the various data tables, and determines
which device is being opened, the code forks. If the device is the keyboard
(device 0) or screen (device 3), OPEN returns. If it's the tape deck (device
1), the code continues by searching for or writing a tape header using the
information supplied in the OPEN call. If it's for the RS232 adapter (device
2), execution jumps to SEROPN to continue the opening process. Finally, if
it's an IEEE device (device >=4), then OPEN calls SENDSA to shift out the
secondary address. When SENDSA returns, OPEN returns to the original
caller. If SENDSA returns with the carry flag set (indicating an error
condition) and then returns, execution will ultimately fall through to an
"ILLEGAL DEVICE ERROR" error.

```
F495      ;===============================================================
F495      ; SENDSA - Send secondary address
F495
F495                SENDSA
F495 A5 B9          LDA SECADR       ;get SA
F497 30 2C          BMI SNDSARC      ;$F4C5 less than 0, exit
F499
F499 A4 B7          LDY FNMLEN       ;get filename length
F49B F0 28          BEQ SNDSARC      ;$F4C5 no filename, just exit
F49D
F49D                                 ; prepare to send filename string
F49D A5 BA          LDA CHANNL       ;get device     number and...
```

```
F49F 20 18 EE           JSR ILISTN+1        ;  command it to listen
F4A2 A5 B9              LDA SECADR          ;get SA
F4A4 09 F0              ORA #%11110000      ;$F0 make it into listen command
F4A6 20 C0 EE           JSR ISECND          ;sent it to IEEE bus
F4A9 A5 90              LDA CSTAT           ;check status variable
F4AB 10 05              BPL SENDSA1         ;$F4B2 OK, then continue
F4AD
F4AD 68                 PLA                 ;error, set stack for caller's caller
F4AE 68                 PLA
F4AF 4C 8A F7           JMP IOERMS5         ;emit "DEVICE NOT PRESENT" error
F4B2
F4B2            SENDSA1                      ; continue processing-send filename
F4B2 A5 B7              LDA FNMLEN          ;get filename length
F4B4 F0 0C              BEQ SNDSARU         ;length is 0, so send unlisten command
F4B6                               ; There is a filename, so send it to IEEE
F4B6 A0 00              LDY #$00            ; set loop counter
F4B8
F4B8            SENDSALP                     ; output filename to IEEE bus
F4B8 B1 BB              LDA (FNPTR),Y       ;get character
F4BA 20 E4 EE           JSR ICIOUT          ;send it
F4BD C8                 INY                 ;get next one
F4BE C4 B7              CPY FNMLEN          ; at the end?
F4C0 D0 F6              BNE SENDSALP        ;no, then loop
F4C2
F4C2            SNDSARU                      ;done sending filename, so send
F4C2 20 04 EF           JSR IUNLSN          ; unlisten command to IEEE bus
F4C5
F4C5            SNDSARC
F4C5 18                 CLC
F4C6 60                 RTS
```

        SENDSA is a simple routine that checks for a valid secondary address
and whether the OPEN command involves a filename. Then, the routine commands
the specified device to "listen" and sends the secondary address and the
filename to it. When completed, the code sends the "unlisten" command to the
device and returns success. If there is a problem sending data to the device,
a "device not present" error is generated.

        The next routine is SEROPN. SEROPN is called by OPEN when it
determines that the device that's being opened is the RS232 port. Here's a
quick story about RS232 support in the VIC.

Based on some information that I've seen on the Web and what's available in
"Inside VIC" and the "VIC20 Programmer's Reference Guide", I've concluded that
the VIC hardware was originally designed to include a MOS 6551 ACIA
communications chip. The 6551, with appropriate level shifters, would have
provided true hardware support for RS232. But for cost reasons, board space
reasons, or both, the 6551 was dropped and emulated in software and in
hardware by using a spare port on one of the existing 6522 VIAs.

The Kernal includes a lot of code to manage RS232 FIFO buffers and to perform
the bit shifting through the User Port (port B of VIA 1). The User Port only
provides TTL-level RS232; Commodore and others sold an adapter that provided
the level-shifting hardware (typically TI 1488/1489 chips) and a DB25F
connector. The circuit is very simple and could have been built by any
resourceful hobbyist.

Without further delay, here's the code that completes the opening process for
the RS232 serial device:

```
F4C7     ;============================================================
F4C7     ; SEROPN - Open RS-232
F4C7     ;  "filename" contains the initialization data for the
F4C7     ;  command and control registers
F4C7
F4C7            SEROPN
F4C7 A9 06              LDA #%00000110      ;set VIA DDR. PB[2:1] are DTR and RTS
F4C9 8D 12 91           STA D1DDRB          ; signals, respectively
F4CC 8D 10 91           STA D1ORB
F4CF A9 EE              LDA #%11101110      ;$EE. Set PCR for CB2/CA2 manual high
F4D1 8D 1C 91           STA D1PCR           ; and CB1/CA1 for H->L IRQ trigger
F4D4                                        ; CB2 is RS232-TX and CB1 is RS232-RX
F4D4 A0 00              LDY #$00
F4D6 8C 97 02           STY RSSTAT          ; clear status byte
F4D9
F4D9            SEROPLP
F4D9 C4 B7              CPY FNMLEN          ;is the filename length = 0?
F4DB F0 0A              BEQ SEROPN1         ;$F4E7 yes, go straight to open
F4DD
F4DD B1 BB              LDA (FNPTR),Y       ;copy first 4 chars of filename...
F4DF 99 93 02           STA M51CTR,Y        ; to buffer and loop
```

```
F4E2 C8                      INY
F4E3 C0 04                   CPY #$04
F4E5 D0 F2                   BNE SEROPLP      ;$F4D9 loop
F4E7
F4E7              SEROPN1                     ;done copying init_data
F4E7 20 27 F0                JSR BITCNT       ;get number of data bits
F4EA 8E 98 02                STX BITNUM       ;save data bits count
F4ED AD 93 02                LDA M51CTR       ;get control register
F4F0 29 0F                   AND #%00001111   ;$0F isolate baud rate bits
F4F2 D0 00                   BNE $+2          ;F4F4
F4F4
F4F4                                          ;convert baud rate bitmap to clock
F4F4                                          ; divisor
F4F4 0A                      ASL A            ;*2
F4F5 AA                      TAX
F4F6 BD 5A FF                LDA R232TB-2,X   ;$FF5A,X baud rate H
F4F9 0A                      ASL A
F4FA A8                      TAY
F4FB BD 5B FF                LDA R232TB-1,X   ;$FF5B,X baud rate L
F4FE 2A                      ROL A
F4FF 48                      PHA
F500 98                      TYA
F501 69 C8                   ADC #$C8
F503 8D 99 02                STA BAUDOF       ;save baud rate divisor
F506 68                      PLA
F507 69 00                   ADC #$00
F509 8D 9A 02                STA BAUDOF+1
F50C
F50C AD 94 02                LDA M51CDR       ; command register
F50F 4A                      LSR A
F510 90 09                   BCC SEROPN2      ;$F51B
F512
F512 AD 20 91                LDA D2ORB        ;check DSR (data set ready)
F515 0A                      ASL A
F516 B0 03                   BCS SEROPN2      ;$F51B ready, continue
F518 4C 16 F0                JMP DSRERR       ;not ready, DSR error
F51B
F51B              SEROPN2                     ; device ready
F51B AD 9B 02                LDA RIDBE        ;set RS232 buffer pointer
F51E 8D 9C 02                STA RIDBS        ; -- RX --
F521 AD 9E 02                LDA RODBE
F524 8D 9D 02                STA RODBS        ; -- TX --
F527
F527                                          ; prepare to "hide" buffer in memory
F527 20 75 FE                JSR IMEMTP+2     ; get MEMTOP
F52A A5 F8                   LDA RIBUF+1      ;has RX buffer been created?
F52C D0 05                   BNE SEROPN3      ;$F533 yes, check on TX buffer
F52E 88                      DEY              ; else create input buffer
F52F 84 F8                   STY RIBUF+1
F531 86 F7                   STX RIBUF
F533
F533              SEROPN3
F533 A5 FA                   LDA ROBUF+1      ;TX buffer created?
F535 D0 05                   BNE SEROPN4      ;$F53C yes, skip create
F537 88                      DEY
F538 84 FA                   STY ROBUF+1      ;else create TX buffer
F53A 86 F9                   STX ROBUF
F53C
F53C              SEROPN4                     ; set MEMTOP to hide buffers from BASIC
F53C 38                      SEC
F53D A9 F0                   LDA #$F0
F53F 4C 7B FE                JMP STOTOP       ;$FE7B set new MEMTOP
```

SEROPN begins by extracting the values of the command and control registers
from the "filename" string passed to it from the OPEN routine. SEROPN then
saves the number of data bits contained in a serial data frame and then
calculates the proper clock divisors for the selected baud rates. The 6551
transmit and receive clocks are emulated by using one of the timers in VIA1 to
shift out the data at the specific bit rate. Receiving bits is accomplished
through the NMI routine discussed in Part 2 of this series.

In a nifty trick, SEROPN allocates transmit and receive buffer memory at the
top of BASIC RAM and lowers the top of RAM variable to fake BASIC into
thinking that there's 512 bytes less of memory.

The SEROPN routine finally exits and returns to OPEN by a JMP to STOTOP, the
routine used to set the top of system memory.

Well, that's all that we have time for. Next time, we'll tackle some of the
tape routines related to the OPEN command.
....

```
....
 ..
 .                                        C=H #20

:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
```

Mods and digital mixing
---------------------- by Jolse Maginnis
                       jmaginni@postoffice.utas.edu.au

As most of you know, MOD playing on the C64 is nothing new, Nate/DAC wrote
modplay64 quite a while ago now, but there has never been any in depth look at
how a mod is played. So in this article I will discuss everything you need to
know about MODs and digital sound mixing (well the main bits anyway!)

Very recently I got sick of working on OS stuff and started working on my own
modplayer for JOS, called, wait for it... Josmod! I was really quite surprised
at the quality of sound it produced, particularly as it was only playing with
4 Bit Mono SID digis.

At that stage Josmod only played 4-channel Amiga format MOD's which have been
around for a long time but are now being outdated by newer formats such as
S3M, XM and IT. These new formats are in the same style as the old Amiga MOD
formats, except they allow for more channels, more instruments/samples and
different effects. Since the SCPU is a fast enough machine to mix and play
more than 4 channels at once, and I was sick of seeing so many MOD's that it
couldn't play because they were done in the newer formats, I added support for
S3M's and more recently, XM mods.

Josmod didn't take very much time to write at all, for one reason in
particular: it's written in C. Compared to writing an application in assembly,
C is so much easier to debug and test. That's one of the great things about
the 65816, you can write in a high level language and get away with it. One
part of Josmod had to be written in 65816, and that's the mixer, which has to
be optimized as it's where 90% of Josmod's time is spent.

Enough propaganda! On with the show..

There are 3 main parts to any modplayer: the loader, the player routine & the
mixer.

First of all I will start by talking about the mixer as it's the simplest part
and also the part that deals with the actual sounds themselves.

The Mixer

First of all, I'd like to point out that I'm no expert on digital sound, and I
only learned all this stuff myself when I started writing Josmod. Here's an
ASCII drawing of a digitally sampled sound:

```
127

       --                                    --
      /  \                   --             /    \
     /    \                 /  \           /      \
 0-----------------------------------------------------------
            \             /        \   /
             \           /          --
              \       /
               ---  /
-128
```
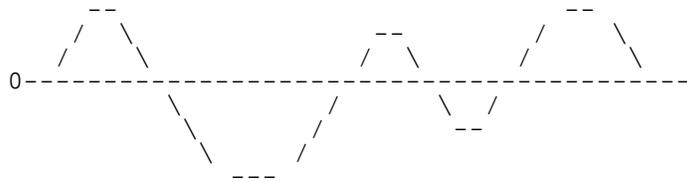
Yes I know it looks crap, but you get the idea. You'll note that it's a SIGNED
sample, which is very important when it comes to mixing.

The Amiga has 4 digital sound channels, each capable of playing its own set of
sampled data, so it can effective play four different sounds at the same time.
But what about the C64? SID's only capable of playing one digi at a time. How
on earth can we get it to play 2 samples at once, or 4? 8 or 16? C64's aren't
alone in their single channel output problem, most PC sound cards only have 1
channel for output (2 if you count Stereo). So how is it done? It's extremely
simple actually. You just add them together! But remember it's signed
addition!

The one thing that you have to worry about when adding though, is that you
don't overflow your result. For example, if you're adding two 8 bit numbers

together you could easily go higher than 127 or lower than -128, just by
having two loud samples. So it's best to use a 16 bit result, and handily the
65816 has 16-bit registers. :) On a side note, if a sample is mixed with an
exact opposite of itself, it turns to silence! I believe this is actually used
in some new cars to cut down on noise inside the car! [Editor's note: it is
also used in things like aviation headsets, and goes under terms like "Active
Noise Reduction" and "Active Noise Cancellation"]

Not only does the mixer have to mix samples together, it also has to control a
couple of other things, namely volume and pitch. The player reads the notes
and effects, and tells the mixer which samples to play, at what speed and how
loud they are supposed to be.

I'll start with volume, as it's quite an easy transformation. In Amiga MODs
volumes range from 0 (Silence) to 64 (Full volume). So a volume of 32 would be
half volume, and effectively divide the samples by 2. e.g. A sample of -40
would become -20 at volume 32, and a sample of 100 would become 50. The best
and fastest way to perform volume calculations in real time is of course
using a lookup table. So you end up with mixer code that looks like this:
(65816 remember!)

```
        lda [Samp]      ; Get the current sample
        and #$ff        ; It's 8 bits only
        asl             ; Multiply by 2 since the volume
        tay             ; table is 16 bit values
        lda [VolTab],y  ; Get the value!
```

VolTab contains a pointer to the appropriate volume table, which is calculated
with:

Sample*Volume/64

Volume was easy but what about pitch? Ordinarily if we were dealing with
normal 1 channel digis we'd simply change our CIA timer rate so data would
be fetched at a different speed, which is ideal. This is basically what the
Amiga does, except it's done in hardware, so doesn't steal valuable CPU
cycles. However, we can't do this as we're mixing various samples into one! So
instead we have to simulate different pitches by altering the sample position
at varying rates, rather than just going straight to the next sample.

So for example, if a sample needs to be played at twice its normal speed, we
add 2 to the sample pointer rather than 1, and it will sound twice as
fast. It's all very well adding values like 1 and 2, but what if the sample
needs to be 1.5 times the normal speed? It's easy solved with some fixed point
math (16 bits for the fraction, 16 bits for the integer part), and you end up
with something like this (remember that Samp is a pointer into the sample
data):

```
        txa             ; Add the fraction part
        adc Low         ; X is used to hold the low part
        tax
        lda Samp        ; Now add the integer with carry
        adc Hi
        sta Samp
        bcc noinc       ; Crossed a bank boundary?
        inc Samp+2
noinc   ....
```

Very simple really isn't it?

Note that the tradeoff here is between playback rate and sample rate; that is,
if we have a sound sampled at 10000 samples per second, and take every other
sample, that's like taking 5000 samples per second. So instead of changing
the _playback_ rate -- which is what e.g. changing the CIA timers does -- we
instead effectively change the original _sample_ rate. Another thing worth
mentioning is that when the playback rate is greater than the sample rate, you
can improve output by "interpolating" the sample. Interpolating is basically
guessing what the sample would have sounded like if it were sampled at the
playback rate. I won't go into it any more than that, as adding it will slow
the mixer down. I believe Nate's latest modplay64 does interpolation.

That's basically all that the mixer does, but I've left out one part, which is
the part that adds the samples together. Rather than adding 4 values together
at a time the Josmod mixer uses a buffer to add all the values for 1 channel
in one go, and then adds all the samples for the next channel to that buffer,
etc. This means that the mixer can handle as many channels as it wants, rather
than being fixed to 4 channels. Here is the code for mixing 2 samples: (Assume
16 bit registers!)

```
        lda [Samp]
        and #$ff
        asl
        tay
        lda [VolTab],y
        adc (OutBuf)            ; Add it to the buffer
        sta (OutBuf)
        txa
        adc Low
        tax
        lda Samp
        adc Hi
        sta Samp
        bcc ninc
        inc Samp+2
ninc    lda [Samp]
        and #$ff
        asl
        tay
        lda [VolTab],y
        ldy #2                  ; Next buffer position
        adc (OutBuf),y
        sta (OutBuf),y
        txa
        adc Low
        tax
        lda Samp
        adc Hi
        sta Samp
        bcc ninc2
        inc Samp+2
ninc2   ...
```

By changing Samp (and Low/Hi) the same code can add different samples to the
buffer. Josmod uses an unrolled mixing loop which does 16 samples before
looping, otherwise there'd be too much loop overhead.

You will note that at the end of mixing all the channels what we're left with
is a buffer that is full of 16 bit samples. Well, I believe the IDE64 team
have a made a 16 bit sound card... but generally this buffer is not going to
be of any use to us, until we convert it to 8 bit unsigned data. The SID uses
4 bit unsigned samples, and the DigiMax and other user port sound cards, use 8
bit unsigned, so it's the best format to convert it to.

Not only does it need to be converted to 8 bit unsigned, but remember the
overflow problem? Values that are too loud need to be clipped to their maximum
volume (127 or -128). So we can kill two birds with one stone here, and create
a post processing lookup table that converts the data and also clips loud
samples. If it were graphed the table would look something like this:

```
$ff -------\
       ^     \
 Clipped      \
               \
                \
$80 ------------------------
                 \
                  \
                   \
                    \
0                    --------
                     ^
                  Clipped
```
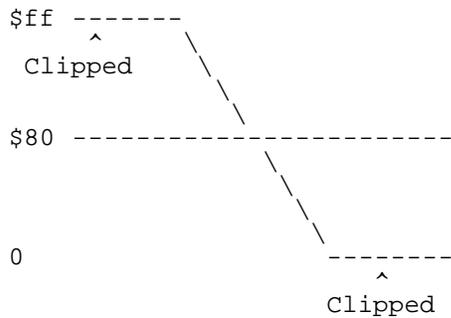
So after all channels are mixed together, the buffer is postprocessed with
that table to prepare it for output on the sound device! Which incidentally
could be any device, as Josmod is device independent, it's up to the device
driver to deal with getting a CIA interrupt and what to do with the
samples. In the case of 4-bit SID this would be:

```
        lda [Samp]
        lsr
        lsr
        lsr
        lsr
        sta @$d418      ; @ means long addressing - $00d418 SID Volume register!
```

That's about all for the mixer, the only other thing that it needs to do is to
check if a sample has ended, and either stop it, or loop it back to some
point.

Bear in mind that the mixer I've described is a mono one, as all channels get
mixed into the one output buffer. A stereo one would mix channels into two
different output buffers.

I should mention that since I wrote this article, I've made some improvements
to the mixer. Namely pre-calculating the sample pointer offset, so that the
mixer loop doesn't need to calculate the next sample position for every
sample. So now the sample mixing code is much quicker:

```
        lda [Samp]
        and #$ff
        asl
        tay
        lda [VolTab],y
        adc (OutBuf)            ; Add it to the buffer
        sta (OutBuf)

        ldy #1                  ; The offset to the next sample
        lda [Samp],y            ; (precalculated)
        and #$ff
        asl
        tay
        lda [VolTab],y
        ldy #2                  ; Next buffer position
        adc (OutBuf),y
        sta (OutBuf),y

        ....

        txa                     ; Add sample speed * 16
        adc Low                 ; This code now only gets
        tax                     ; executed once per 16
        lda Samp                ; samples rather than every
        adc Hi                  ; sample.
        sta Samp
        bcc ninc2
        inc Samp+2
ninc2   ...
```

These changes provide, at a guess, a dramatic increase of about 30 percent in
playback speed. The only drawback to this method is very slight loss of
accuracy, but I don't think you will be able to tell the difference.

The Loader

You've now seen how to mix samples, so what you need to know next is how to
load in the samples from the modfile and also load the data that tells you how
to play them. Ok let's check out the .mod file format. Sizes are all in bytes.

Offs : Size
----
0   : 20   : Module name
20   : 930 : Sample headers for 31 Samples
950 : 1   : Number of orders
951 : 1   : Unused?
952 : 128 : The orders
1080 : 4   : Identification string
1084 : ?   : The patterns
?   : ?   : The 8 bit sample data for all samples

Ok first of all you need to know exactly what orders & patterns are.

Patterns contain all the information about what pitch and sample to play and
also contain information about effects which change the volume and pitch of
samples. Every pattern contains 64 rows of data. Each row holds note
information for the number of channels contained in the tune (normally 4 in
Mods). A pattern can be visualised as something like this:

```
          1         2         3         4
------------------------------------
1:  │ 1 c-1 │ 2 d-2 │ 2 f#5 │ 3 e-4 │
2:  │ 5 d-1 │ - --- │ - --- │ 3 f-4 │
......
64: │ 4 e-1 │ 3 d-1 │ - --- │ - --- │
------------------------------------
```

That's a simplified version of what a pattern is like, as it doesn't include
any effect information.

So what does "1 c-1" mean?

Well the first 1 refers to the sample number, while "c-1" refers to the note
and octave to play (C natural, octave 1).

Orders will be familiar to anyone who's had any experience with some of the
SID editors/players. Rather than playing pattern 0, pattern 1, etc.. through
to the last pattern, orders allow you to choose the order in which patterns
are played, thus allowing you to repeat patterns easily. E.g. The order table
may look something like this: 0 1 2 3 2 2 4 3 5 6

You will notice that nowhere in the mod format does it mention how many
patterns are contained in the file.. To calculate this you look through the
order table and find the highest value, and use that as the number of patterns
in the file (+1).

Ok now let's have a look at the sample header format:

```
0     : 22   : Name
22    : 2    : Length (divided by 2)
24    : 1    : Finetune
25    : 1    : Default Volume (0-64)
26    : 2    : Loop position (divided by 2)
28    : 2    : Loop length (divided by 2)
30
```

One thing to be warey of here is the fact that the 16 bit values stored in
this structure are in big endian format, because Amigas use the 680x0 line of
processors. So to convert it for use with C64 it needs the high and low bytes
swapped around. Also I have no idea why the 16 bit values are stored divided
by 2... I guess this is something to do with playing the mod on the Amiga too,
or perhaps to allow larger than 64k samples?

Ok now we know what the format of a .mod file is, it's time to get it loaded
so our player can play it!

You will notice that the first 1084 bytes of the file are fixed, so one
approach to loading would be to load those first 1084 bytes and go on from
there, but the approach I use is to load each part of the header seperately,
which makes it easier to put the data in a format easier for the player. Ok
let's take a look at the C source for Josmod's .mod loader..

First a couple of declaration for identifying which mod format it is:

```c
typedef struct IdentStruct {
        char *String;
        int channels;
} Ident;

static Ident idents[] = {
        {"2CHN", 2},
        {"M.K.", 4},
        {"M!K!", 4},
        {"FLT4", 4},
        {"4CHN", 4},
        {"6CHN", 6},
        {"8CHN", 8},
        {"CD81", 8}
};

int loadMod(char *name, ModHead *mp) {

        /* Variable declarations */

        FILE *fp;
        S3MSamp *samp;
        unsigned char *patp;
        unsigned int i,j,temp,numpat,temp2;
        long patsize;
        char *samdata;
        Ident *idp;

        /* Open the mod */

        fp = fopen(name, "rb");
        if (!fp) {
                perror("josmod");
                exit(1);
        }

        /* Read the name of the mod */
```

```c
        fread(mp->Name,1,20,fp);
        mp->Name[20]=0;              // C strings are null terminated!

        /* Read 31 sample headers */

        samp = &mp->Samples[0];
        for (i=0;i<31;i++) {

                /* Read name and null terminate */

                fread(samp->Name,1,22,fp);
                samp->Name[22]=0;


                /* Read length
                Note: readWord() converts the 16 bit value
                into C64 format
                */

                samp->Length = readWord(fp) << 1;

                /* Read and calculate finetune
                Finetune is a 4 bit signed value
                Values over 8 are negative values
                */

                samp->Finetune = fgetc(fp);
                if (samp->Finetune>=8)
                        samp->Finetune -= 16;

                /* Read volume */

                samp->Volume = fgetc(fp);

                /* Read and calculate ending/looping positions */

                temp = readWord(fp) << 1;
                samp->Replen = readWord(fp) << 1;

                /* If the loop length is higher than 2, this
                is a looping sample, otherwise it's a standard
                non looping sample that stops at the end */

                if (samp->Replen > 2) {
                        samp->Looped = 1;
                        samp->End = (char *) temp + samp->Replen;
                        if ((unsigned int) samp->End > samp->Length)
                                samp->End = (char *) samp->Length;
                } else {
                        samp->End = (char *) samp->Length;
                        samp->Looped = 0;
                }

                /* Go to next sample structure */
                samp++;
        }

        /* Read orders and calculate number of patterns */

        mp->NumOrders = fgetc(fp); fgetc(fp);
        numpat=0;
        for (i=0;i<128;i++) {
                temp = fgetc(fp);
                if (temp > numpat)
                        numpat = temp;
                mp->Orders[i] = temp;
        }
        numpat++;
        mp->NumPatterns = numpat;

        /* Identify type of mod */

        fread(mp->Type,1,4,fp);
        mp->Type[4]=0;

        mp->Channels=0;
        idp = &idents[0];
        for (i=0;i<8;i++) {
                if (!strncmp(idp->String,mp->Type,4)) {
                        mp->Channels=idp->channels;
```

```
                                continue;
                        }
                        idp++;
                }

        /* If we couldn't identify it, this isn't a mod file!
        close the file and return. Otherwise print number of
        channels. */

        if (!mp->Channels) {
                fclose(fp);
                return 0;
        } else {
                printf("Channels %d\n",mp->Channels);
        }


        /* Calculate pattern sizes and allocate memory for patterns */

        mp->LineSize = 4 * mp->Channels;
        mp->PatSize = 64 * mp->LineSize;
        patsize = mp->PatSize * numpat;
        patp = xmalloc(patsize);
        mp->Patterns = patp;

        /* Load patterns */

        for (i=0;i<numpat;i++) {
                for (j=mp->PatSize/4;j;j--) {
                        /* Each note is stored in 4 bytes.
                        Contained in the 4 bytes is:

                        Sample number.
                        Note
                        Effect
                        Effect Parameter

                        conv2Note() converts the mod format note format
                        into a format easier for Josmod.

                        */

                        * (unsigned int *) (patp+2) = readWord(fp);
                        * (unsigned int *) patp = readWord(fp);
                        conv2Note(patp);
                        patp += 4;
                }
        }

        /* Load Samples */

        samp = &mp->Samples[0];
        for (i=0;i<31;i++) {
                printf("Sample %d: %s\n",i,samp->Name);
                if (samp->Length) {

                        /* If the sample exists, allocate memory
                        for it, and read it in. Then fix up it's end
                        pointer, then call fixSamp() to prepare it for
                        mixer output. */

                        samdata = xmalloc((long) samp->Length + 128);
                        fread(samdata,1,samp->Length,fp);
                        samp->Samp = samdata;
                        samp->End = (unsigned long) samp->End + samdata;
                        fixSamp(samp->Looped, samp->End, samp->Replen, 128);

                } else samp->Samp = NULL;
                samp++;
        }

        /* Close the file and return successfully */

        fclose(fp);
        return  1;
}
```

That's all for the loader! Now that patterns, orders, and samples are all
loaded and in a suitable format, the player is ready to play it!

Unfortunately that's all for now, as the player is quite a complicated topic,

I'll save it for a later article. But before I finish, I do want to give you a
basic outline of what happens in the player, so you can see where the mixer
fits into all this.

The Player

Mods have two speed variables which control the way a mod is played. One is
BPM, or Beats Per Minute. BPM specifies the speed at which the mod player
routine is called, the default value is 125 BPM, or 50 times a second. As we
all know, 50hz is the speed of PAL TV's, so it's no coincidence that this is
the default, as it allowed Amiga players to setup their play routines on an
interrupt synchronised with the screen, the same as with SID players.

BUT since we don't have the luxury of having hardware to play the samples for
us, it's not a good idea to hook our player to a raster interrupt, instead we
calculate how many output samples need to mixed before we update our mod. We
calculate that with the following:

playbackrate / (BPM*2 / 5)

This value is what I call the "TickSize". Everytime the player is updated, it
will mix TickSize bytes of samples, then loop and update the mod again, mix
TickSize bytes again, etc..

The other speed variable is the number of Ticks per beat, and defaults to 6.
This means that the mod player only fetches the next row of pattern data every
6 Ticks. E.g. After 6*TickSize samples have been mixed.

Each channel in the Mod has assocatied with it, a mixer structure, which looks
like this:

```
typedef struct {
        char *Samp;                 // Current sample position
        char *End;                  // End or loop position
        unsigned long Replen;       // How far to loop back
        unsigned long Speed;        // Speed/pitch of sample
        int Low;                    // Fractional part of sample position
        int Repeats;                // Whether or not it repeats
        int *VolTab;                // Pointer to volume table
        int Active;                 // Whether to mix this channel
                                    // or not
} MixChan;
```

The player has to prepare these structures for each channel and on every tick
update the appropriate fields depending on the data contained in the patterns.

I'll leave you with psuedo code for what the player looks like:

```
tickbeat = 6
row = 64
tickcount = 0
orderup = 0
make all channels inactive
while (orderup < numorders) {
        if (row == 64) {
                get next pattern from orders
                orderup++
        }
        if (tickcount == 0) {
                get row data from pattern
                process note - set sample speed
                process sample number - set sample position volume
                process effect - speed, volume and position effects
        } else {
                process effect
        }
        mix ticksize samples
        if (mixing buffer full) {
                play buffer
        }
        tickcount++
        if (tickcount >= tickbeat) {
                tickcount = 0
                move to next row
                row++
        }
}
```

That's basically what the mixer does. The majority of the player code deals
with effects processing, most of which are easy to process. Effects include:

```
Portamento - sliding a note's pitch up or down.
Volume setting/sliding
BPM and Tick Beat changes
Pattern looping
Vibrato and Tremolo - Slight alterations in pitch and volume.
```

Some effects are only processed on Tick 0, and some are processed on other ticks.

Anyway that's all for now, the player code will be examined in a later article. In the meantime you can view the source code in HTML form at: http://jos64.com/src/josmod.c.html
```
.......
....
..
.                                   C=H #20
```

:::

```
<=============
  The C64 Digi
  =============> Robin Harbron <macbeth@psw.ca>
                 Levente Harsfalvi <levente@terrasoft.hu>
                  Stephen Judd <sjudd@ffd2.com>
```

Introduction

Digis -- digitally sampled audio -- are fairly common on the 64. This is meant to be a comprehensive article on digis: how they work, examples, different playback methods on the 64 (volume register and Pulse Width Modulation), and some tricks. We'll even show you how to play 6-bit and even 8-bit digis in high quality on a 64, which is really pretty neat to hear.

The first part discusses digis from a fundamental point of view -- just what a digi is, acoustic signals, and things like that. The most common method of playing digis is via the volume register at $d418, and the next two sections are devoted to this technique. Section two discusses some SID fundamentals, and the reason why $d418 may be used for digis (and why later-model SIDs don't play digis correctly); Section three discusses $d418-digis from a software perspective: how to play them, tricks for improving them, how to boost digis on 8580 SIDs, and how to detect what kind of SID (6581 or 8580) is in the machine. The fourth and final part of this article discusses pulse width modulation, and includes example source code and a binary that plays a true 7-bit digi at around 16KHz -- something which, we think, has never been done before.

Without further ado...

===============
Digis: Overview
===============

        The whole point of playing a digi on a 64 is to provide something for your ear to hear. So let's begin by discussing just what an acoustic signal is and how that relates to digis.

        Probably everyone knows that "sound" is how your ear responds to changes in air pressure -- that is, when you clap your hands together, it compresses the air between your hands in a special way, and that higher pressure moves outwards into the surrounding air (since it's at lower pressure). That pressure change propagates along and when it encounters your ear it causes the ear drums to move, causing three little bones to move, causing some fluid to move, causing tiny, exquisitely sensitive hairs to move, transmitting a signal that your brain converts to "sound".

        An audio speaker also changes the air pressure in response to a signal. If you take a coil of wire and change the voltage on it, it generates a magnetic field; if a magnet is placed inside the coil, the changing magnetic field will place a force on the magnet, causing it to move, causing some air to be pushed along, causing a change in pressure, causing a signal to propagate to your ear which your brain interprets as Van Halen. All a stereo (CD player, etc.) does is send a varying voltage signal to the speaker. As that voltage level goes up and down the magnet moves back and forth, and so the speaker converts that electrical energy into an accoustic wave.

        For us, the trick is to coax SID into sending a specific voltage

signal to the speaker, the way a stereo or CD player might. And a CD player
is of course a very apt comparison, since it is itself a digi player.

        Just for reference, a really good pair of ears can hear signals from
around 20Hz to 22KHz, with the sensitivity dropping considerably outside
of around 100Hz to 10KHz. A CD player has a playback rate of 44KHz, and
the highest frequency SID can generate from the frequency registers is
around 4KHz. If you've ever set SID to maximum frequency and heard just
how high 4KHz is, you can appreciate that even 10KHz is _really_ high, and
actually quite difficult to hear. In human speech, most of the information
content of vowel sounds is contained in the range 300Hz - 3KHz, and above
around 1KHz for consonant sounds; most information in musical sounds is in
the range 100Hz - 3KHz.

Discrete Sampling

        To understand digis a little better, consider the more general
case of a discretely sampled signal -- a continuous signal sampled at
discrete time intervals. Let's say we had some device producing a
continuous sinusoidal signal in time:

          *   *
        *   *
      *   *
     *   *
    *   *
   *   *   *
  *   *   *
                     *   *   *
                      *   *
                       *   *
                        *   *
                         *   *   *
                          *   *
--> time

(yes, I did miss my calling as an ASCII artist)

To turn the signal into a _discrete_ signal, we simply sample the
signal at discrete intervals of time. For example, let's say the above
signal lasts one second, and is input into a device which measures the
value every 1/4-second. The device will spit out four numbers: 0, 1, 0,
and -1:

            *

*   *

            *

The sampling frequency here is four samples per second -- 4 Hz. If we were
to then play back this signal at the sampling frequency, we'd get a signal
like

        ****

**********   **********

          ****

So one thing sampling does is to "staircase" a signal -- the sample becomes
some sort of "average" value over the sample period. Increasing the sample
rate -- taking more samples per second -- will smooth things out, and the
sampled signal will look (and sound!) more like the original signal.

Now let's say we just took two samples in that one second -- 2 Hz sampling
rate -- and just happened to catch the signal at its maximum and minimum
values (the peak and trough). Upon playback, the signal would look like

****
                    *
                    *
                    *
                    *
                    *
                    *

```
*********************
```

That is, a square (pulse) wave. If you're on the ball, you've noticed
that the frequency of the new signal is 1 Hz -- exactly half the sampling
frequency. This is also called the Nyquist frequency. In general, the
maximum frequency that can be captured in a discrete sample (called the
Nyquist critical frequency) is half the sampling frequency -- as you can
see above, it takes two data points to get a single (nonzero) frequency.
So, for example, the highest frequency a CD player -- which has a sampling/
playback rate of 44KHz -- can capture is 22KHz, well above the range of
normal human hearing.

Thus, increasing the sample rate increases the frequency range captured
in the discrete signal. This is why a digi at a high sample rate in general
sounds better than a digi sampled at a low sample rate.

BUT -- there is more to life than sample rate: there is also sample
resolution. The sample resolution -- 4-bit samples, 8-bit samples, etc. --
determines how accurately the sample measures the actual signal. For
example, let's say we sample sin(x) when x=0.5:

        sin(0.5) = 0.4794255...

No matter what sample resolution we use, there will always be some error
in the measurement, and the _true_ value of the sample will be the
measured value plus some error.

In general the sampling errors are random and uniformly distributed, so
the sampled signal corresponds to the original signal plus some noise (the
random errors). That is why you almost always hear some sort of hiss on
a normal C64 digi, which uses a resolution of 4 bits per sample.

So, increasing the sample _resolution_ decreases the amount of noise introduced
into the sampled signal (and increases the dynamic range), and increasing the
sample _rate_ increases the frequency range.

If you're _really_ on the ball, you've noticed that the 1-Hz square pulse
above actually contains frequencies higher than 1Hz, simply because a
square pulse contains higher harmonics in addition to the 1Hz fundamental
frequency. And you've also no doubt realized that the sampled pulse wave
would sound different than the original sine wave (due, of course, to the
added harmonics) -- it's at the right frequency, but it will sound like a
pulse wave instead of a sinusoid.

Have we somehow broken the Nyquist limit?

The answer is no, because of a nifty thing called the Discrete Sampling
Theorem, which says that, given the samples h_n of a bandwidth-limited
function h(t), the original function h(t) is given by

        h(t) = dt * Sum{ h_n * sin(2*pi*f_c*(t-n*dt)) / (pi*(t-n*dt)) }

where dt is the sampling period and f_c is the cutoff/critical frequency.

What this means is that the original signal can be _reconstructed_ from the
discrete samples, not that it is _equivalent_ to the discrete samples.
The Nyquist limit is the highest frequency that can be _reconstructed_ from
the discrete samples, not the highest frequency that will be produced if you
"staircase" the discrete samples through a speaker. If the original
signal is bandwidth-limited, and there are at least two samples for the
highest frequency, then the signal can be completely reconstructed.

Since a "normal" digi contains all these extra frequencies, shouldn't a digi
sound "different" than a "true" analog signal? Sure. On the other hand, many
of the extra frequencies are beyond the range of human hearing, and the rest
can often be removed using a filter -- all CD players filter the output, for
example. So sometimes it is worthwhile to turn on a low/band pass filter
when playing a C64 digi, especially at lower sample rates.

And that more or less summarizes basic discrete sampling theory.


```
=============
D418 Playback -- Hardware
=============
```

The SID contains both analog and digital subparts on one silicon plate -- in
other words, it is a mixed signal device.

At the time, the SID was certainly the best of the microcomputer sound chips.

This may be mostly due to its mixed signal design, which the designers used to solve certain problems.

The hard thing in a sound generator design is to implement waveforms, volume control, and mixing. Things like that don't really fit into the digital 'either 0 or 1' philosophy, unless lot of data bits and arithmetic functions are involved. In a fully digital sound chip, the waveforms could be generated by ROM lookup tables. The mixing function could be derived from binary addition, while the volume control from division or multiplication. Unless the sound functionality is greatly simplified, the arithmetic functions must be present and they must be implemented in hardware. Finally, the D/A conversion could be done by (fast) pulse width modulation just at the output stage. (Most of today's wavetable sound cards operate like this).

This method implies heavy arithmetic hardware, which was not an option for designers back then. Still, most sound chips were fully digital, and all suffer from the required compromises (i.e. generating square waves only, no dedicated channel volume control, etc. - both TED and the VIC-I are obvious examples).

The solution that one finds in the SID design is very straightforward: mixing and variable volume level is problematic in a digital circuit when dealing with waveforms, so simply avoid doing it. In the SID, only the microcomputer interface, the registers, the oscillators (phase accumulating oscillators), and other controller logic are digital; the mixing and volume control parts are fully analog. There are digital to analog converters providing analog voltage levels from the digital state variables. The SID D/As are in fact 'multiplying' D/As, having an analog input (AIN), an input base voltage (IBASE), and a digital input. They operate by amplifying the input voltage offset (AIN-IBASE) by a factor proportional to the number on the digital input and adding this offset back to the base level.

This mixed signal design also allowed some other features to be implemented. The most important one is the analog filter (that is, a two integrator loop, bi-quadratic filter, according to Yannes). With that, the SID points beyond a home computer sound chip - it is a true analog subtractive synth (marketing as such was cancelled because of manufacturing capacity reasons).

Here is a detailed map on the SID inners (analog path; probably my most beautiful ASCII ever :-D). Info can be found in the SID patents (US 4,677,890; 1986), the MOS 6581 technical document (can be found somewhere on the Net), or the back of the Programmer's Reference Guide (PRG).

```
                  ------------------   11bit  ------------
                  |Cutoff freq reg|-------->|Cutoff D/A|---------o
                  ------------------         ------------        |
                     $d415-16                                    |
                                                                 |
                  -----------------   4bit   ------------        |
                  |Resonance reg. |-------->|Reson. D/A |-o      |
                  -----------------          ------------  |     |
                     $d417.[4-7]                           |     |
                                                           |     |
                                       =0                  v     v
-----------   -----------     >o------------>|    ------------------
|wave D/A |--->|env. D/A |-->o/             |    |                |
-----------   -----------     o--->|        |    |                |
     ^             ^         ^ =1   o--------|--->|                |
     |12bit        |8bit     |                    |                |
     |             |         |                    |                |
-----------   -----------    |                    |                |
|OSC1 +   |   |ADSR cnt+|    |                    |                |
|wave sel.|   |env. log.|    |                    |                |
-----------   -----------    |                    |                |
$d400-03,     $d405-06,    $d417.0                |                |
$d404.[1-7]   $d404.0                             |    FILTER      |
                                                  |                |
                                       =0         |                |
-----------   -----------     >o----|------->|    |                |
|wave D/A |--->|env. D/A |-->o/             |    |                |
-----------   -----------     o--->|        |    |                |
     ^             ^         ^ =1            |    |                |
     |12bit        |8bit     |               |    |                |
     |             |         |               |    |                |
-----------   -----------    |               |    |                |
|OSC2 +   |   |ADSR cnt+|    |               |    |                |
|wave sel.|   |env. log.|    |               |    |                |
-----------   -----------    |               |    |                |
$d407-0a,     $d40c-0d,    $d417.1           |    |                |
$d40b.[1-7]   $d40b.0                         |    | LP   BP   HP   |
```

```
                               |         =0  |   ------------------
                               |         >o->|   |      |      |
                        =0   >o----|--o/      ***  o      o      o
-----------   -----------   >o---- |--o/       o-        /      /      /
|wave D/A |--->|env. D/A |-->o/     |    o-     =0 V      V      V    =1
-----------   -----------   o--->   |    ^   =1    o  o    o  o    o  o
      ^             ^       ^ =1    |                | |    | |    | |
      |12bit        |8bit   |       $d418.7 |<-------o  |    |      |
      |             |       |               <-----------o    |      |
-----------   -----------   |               <----------------o      |
|OSC3 +   |   |ADSR cnt+|   |
|wave sel.|   |env. log.|   |
-----------   -----------   |
$d40e-11,     $d413-14,   $d417.2  |
$d412.[1-7]   $d412.0            |       ------------------
                                 |       | Master volume |AUDIO
                      =0       o--->| D/A            |----->
                    >o----|------->|   ------------------ OUT
EXT IN -------------------->o/             ^
                    o--->|                 |4bit
                    ^ =1                    |
                    |                   $d418.[0-3]
              $d417.3  ^               |
                    |                  |
         Analog mixing ---|--------|
```

***: Filter type select switches, $d418.[4-6] respectively

$d418 digis

The most common method of playing a digi is to use the register at $d418.
When someone plays a digi using the master volume register, the situation is
similar to the waveform D/A converters. Both D/As are multiplying D/As --
signal amplifiers whose amplification is proportional to the input digital
number. If there is a nonzero signal offset on the D/A input it will be
multiplied proportionally by this number.

Playing digis with $d418 is possible because there is indeed a relatively
large DC voltage offset on the master volume D/A. This offset is present
right from the moment when the SID is powered up.

Where can this DC offset come from?

There is a mixer before the master volume D/A (see figure). If there's a DC
offset on the D/A input, it must come from there. ...And going further,
the DC offset on the mixer must also come from somewhere. But where?

Signals come from the three ADSR volume D/As, the EXTIN line, and the three
outputs of the filter. Fortunately, all paths that go to the mixer have analog
switches (all paths can be disconnected from the mixer individually, if that's
needed).

The above analog switches are driven by the filter selection bits ($d417 bits
0-3), the voice 3 off bit ($d418 bit 7) and the filter type selection bits
($d418 bits 4-6).

After a reset, the filter selector bits are all 0 (all signals are routed
towards the master mixer), the 'voice 3 off' switch is on, and the filter type
selector bits are 0 (filter outputs are unconnected). In this state, only
EXTIN and the three SID voice signals are present on the mixer. EXTIN can
be eliminated as the source since it has no DC offset (as long as the computer
was not hacked, see notes on the 8580).

The ADSR volume D/A is similar to the previously mentioned multiplying D/As.
If the digital number on the input is 0, the input analog signal offset can't
pass through (as measurements verify). This is the case when SID is reset,
setting the envelope counters to zero. Therefore, nothing behind the ADSR
multiplying D/As can have any effect on the DC offset of the mixer.

So, the DC offset must come from the ADSR multiplying D/As. Another
measurement shows that even the mixer itself has a small DC offset.

Tests and results

I did some tests that support this theory. They were done 'by hand', by simply
using a digital voltmeter + the FC3 monitor.

The chip was a 6581(R1), 0883, Hong Kong (an early 6581).

When turned on the voltage on the AUDIO OUT was about 5.5 volts (slowly
decreasing as it warmed up, stopping at about 5.43 after some 10 mins - all
subsequent tests were done after this time period).

Writing $0f to $d418 raised the output voltage to 6.15 volts. Therefore, the
maximum output amplitude that can be achieved when playing digis is 0.72 volts
in this 'mode' (without wiggling any other SID settings to achieve higher
voltage levels) -- remember that what counts is the maximum voltage
difference, not the maximum absolute voltage.

The next test is to determine if the mixer has its own DC offset (with all
possible paths are disconnected). It's possible to do. With the volume at
maximum (to maximize any effect), all voices are routed towards the filter
($d417 = $0f), while making sure that the filter outputs are not routed to the
mixer ($d418 = $0f). In this state no paths can drive the mixer. The result
is 5.39 volts. When the volume changes, the output also changes towards the
previous 5.43 volts --> there is a (very small) DC offset from just the mixer.

What could be the DC offset value of each individual SID voice (i.e. the base
level difference of the multiplying D/As)? Doing the above, but leaving one
voice routed to the mixer ($d417 = $0e, $0d or $0b) gives 5.69 volts.
5.69-5.43 = 0.26 volts, and 5.43 + 3*0.26 = 6.21, almost 6.15 volts.

To determine if the ADSR multiplying D/As act as expected, I used pulse
waves with zero frequency and 0 or $fff pulse width (two cases), to make the
input signal of the ADSR multiplying D/A the minimum and maximum possible
level. After careful checking, the output changed a few hundredth volts
(about 0.01 volt per voice). So the D/A doesn't close up completely, but
it's still O.K.

To prove that these offsets are equal for all voices, I did another test. Some
people know that the filter inverts phase (multiplies the input signal by -1).
Machine is reset, $d417 = $01, $d418 = $9f. (Voice 1 is routed through the
filter, voice 3 is cut off from the mixer completely ($d418.7), low pass
filter is selected, volume = $0f). The output voltage was 5.41 volts, just
very slightly below the "default" output level. This means that the DC of
voice2 + (-1*) DC of voice 1 resulted in about 0 relative offset. Doing
similar tests proved that the DC offsets for the voices match each other
almost exactly (within a few hundredths of a volt).

These measurements all support the idea that the DC offset comes from the
ADSR multiplying D/As, that the offset is mostly independent from the waveform
D/A converters (as long as sustain levels are 0), and that the offsets are
equal for all voices. In addition, a small DC offset is supplied by the master
signal mixer itself.

What if we try different sustain settings? For this test, set the volume to
maximum, as usual. Set the sustain level to $0f for all voices ($d406, $d40d,
$d414 = $f0). Start the attack, but with no waveform selected ($d404, $d40b,
$d412 = 01). The output level is now 5.21 volts, a little bit below the '0'
offset of the audio output! (Doing the test with just one voice (all
others disconnected), the output is 5.29 volts).

Finally, we can do some experiments with the pulse waveform. The pulse
waveform is useful for these tests, since at zero frequency we can set both
the minimum and the maximum constant DC levels at the voice D/A just by using
the pulse width registers. Reset the computer. Set voice 1 to zero frequency,
pulse level $0fff, sustain level 15, and $d404=$41 (pulse waveform + gate on).
Route only voice 1 through the mixer ($d417 = $0e). The output voltage is
similar to the test when no waveform was selected -- 5.29 volts! This seems to
show that "waveform accu = $0fff" is the same as when no waveform is selected
(i.e. the waveform D/A digital input pins are pulled high when they're not
driven, as seen in most other NMOS chips).

When the pulse width is 0 in the above test the output changes to 6.34 volts.
This seems to be strange (a multiplying D/A giving higher signal level for
multiplying something by 0).

Now, when the ADSR multiplying D/A is closed, the output is 5.70 volts. When
it's fully open, the output changes from 5.29 volts (wave acc= $fff) to 6.34
volts (wave acc = 0). One reasonable answer is that the base voltage of the
waveform D/A is higher, and the analog input is tied lower than the base
voltage of the ADSR D/A -- the effect is that the SID waveforms will lie
'around' the ADSR multiplying D/A base voltage, more or less symmetrically.

This was surely done intentionally, to reduce absolute voltage levels (for
linearity). In the 6581, the big DC offset is probably a result of having the
ADSR D/As and the master volume D/A at different base levels (the difference
appears as true DC offset on the master volume D/A). If both were the same
(presumably at VDD/2), and the waveform D/A parameters were selected similarly

(operation is symmetric to VDD/2), there would be no final DC offset at all.
Rather like the 8580...

Other issues

So now we know why $d418 digis are possible - but still, there are some things
to note.

The DC offset on the master volume D/A changes with different SID settings,
and whatever affects the DC offset on the mixer will affect the digi
volume. For example, even the filter output signals have a small DC offset.
Just do a test - set the volume to 0f, then simply turn one filter route on
(for example, $d418 = $1f). You'll hear a small click (i.e. a small DC offset
change on the mixer), even if the filter has no input.

Moreover, as seen above, the DC offset can be eliminated completely (just by
SID register settings), leading to no audible digi sound at the output. In
other words, whatever affects the DC offset on the mixer _will_ affect the
digi volume.

One place where this is important is playing a digi with a tune: there's a
constantly changing signal going to the mixer instead of a constant DC offset,
so playing a digi on the master volume also causes distortion for both the SID
voices and the digi sound (since they're cross modulated). To reduce this
effect most 3+1 like SID + digi players play samples by writing 8-offset sample
values to $d418 (ie. adding 8 to 3-bit sample values and writing this to $d418
- see players used by Jeroen Tel and other famous composers using digi). This
trick reduces the modulating effect while still maintaining good digi volume.

The DC offsets used to create awful clicking sometimes. For example, the
filter inverts phase. If the filter is currently routed to the mixer, there'll
be a large 'click' (2 times the DC offset) when a voice is on and its routing
is changed to or from the filter.

The 8580

This is a completely redesigned chip. I don't know details, but it was
probably redesigned by the time all other chips in the C64 were done for CSGs
new manufacturing technology and the C64c. It is a 'better' chip from the
technical side (but in my opinion it sounds crude in comparison to the 6581,
at least the R4 series). The 6581 was designed in months. Bob Yannes had to do
everything from scratch and use the manufacturing technology MOS currently had
(NMOS). And it shows. First, it has high background noise. The DC offsets are
really also a misfeature. The D/A converters are sometimes non-monotonic (at
least, the waveform D/As and the filter cutoff D/A have some drops at the
change of the most significant bits). The op-amps in the active (resonant)
filter are simple, linearized NMOS inverters ;-) (loopbacked, they act like
more or less linear op-amplifiers around VDD/2). And I still haven't mentioned
bugs in the digital side (ADSR envelope bugs). Because of the above, one
probably won't find two identical 6581 chips -- each sounds a little bit
different (mostly due to the filter). Since the active components of the
filter are far from ideal, the filter is strongly nonlinear (the cutoff curve
changes with signal amplitude). On the other hand, these things are what make
the SID sound so unique.

Most of the problems were fixed in the 8580. It has much less background
noise. The chips sound the same (there are hardly any differences between
different 8580s). Most of the DC offset issues (the clicks) were eliminated. It
needs less power, and lower VDD level. Something was changed also in the
digital logic, but the ADSR part was not touched. The 'combined' waveforms are
a bit different (and more useable from the musician's point of view).

The clicks were reduced, which means that there is no (or no significant) DC
offset on the master volume D/A in the 8580.

(I have not done any measurements, but after listening to a lot of 3+1 channel
type musics, I have a strong suspicion that even if sounds are turned on, the
average DC offset on the master volume D/A is still minimal).

To fix this in software, you'll have to wait until the next section of this
article.

To fix this in hardware, people use a simple hack: take a resistor of about
330k and tie the SID EXTIN line to GND through that (directly, beside the
chip, on the mainboard).

The EXTIN line goes directly to the mixer, and thus the master volume D/A, or

can also be routed through the filter. In either case, unless the filter is
disconnected, the above hack will give a pretty large DC offset, similar to
the original 6581s. So, digi sounds can be played :-) (even with SID music
playing simultaneously, similar to the 6581).

This solution is good as a work around, but there's one thing to note: this is
not completely the same as the 6581 ADSR D/A offset voltage. At least, this
offset is negative (should that pin rather be tied to VDD?). Programs that
depend on the 6581s way of DC offsets will not work correctly (but I know of
very few such programs, so at worst you'll experience slightly different digi
sound only occasionally -- but hey, the 8580 sounds different anyway). Another
problem is that when EXTIN is routed through the filter the DC offset may
cause strong distortion since the DC operating point of the filter is changed
-- bad news if the 'semi-linear' amplifiers in the filter are picky about
absolute DC level. Some music (not neccessarily involved with digis) indeed
do route EXTIN through the filter, for noise reduction on older C64s (with the
earlier C64 mainboards that pick up lots of 'digital' background noise from
EXTIN). DC distortion can also occur occasionally for the same reason but on
the master volume D/A (the higher the difference from VDD/2, the greater the
risk of experiencing nonlinearity and clipping distortion).

Some final words

A lot of this information comes from Dag Lem, who is certainly the No. 1 SID
hacker for me ;-). Take a look at reSID, his SID emulator library (the sources
can be downloaded from somewhere). reSID contains so much reverse engineered
information of the real SID that you won't believe it -- check it out if
you're interested.

=============
D418 Playback -- Software
=============

$D418 digis are by far the most common playback method. The volume register
gives 16 different amplitudes (0-15), and so can provide 4-bit digi playback.

In its most basic form, this is an extremely easy routine to code. Simply
load each 4-bit sample, and store it in the volume register ($d418).
Assuming $fd/$fe are pointing to the beginning of a series of samples, the
following code will play it back:

```
        ldy #0

:loop   lda ($fd),y
        sta $d418

        ldx #5 ;some delay value
:delay  dex
        bne :delay

        iny
        bne :loop

        inc $fe
        jmp :loop
```

The ldx #5 would have to be adjusted depending on the speed of the
sample - the lower this number (not including zero) the faster the
sample will play back.

There are a number of improvements we could make to this code - first
of all, this method takes twice as much RAM to store the sample as is
necessary. Because we're dealing with 4-bit samples, we can store 2
samples in each byte. This can be handled simply by alternately
masking out the high bits (with AND #15) to play the sample stored in
the low nybble, and by shifting the high nybble down to the low nybble
to play the high nybble (LSR : LSR : LSR : LSR). A lookup table may also
be used to save processor cycles (but use more RAM).

Another improvement is to move the routine to zero-page, and use self-
modifying code. In general, this results in the fastest digi players.

We should of course have the routine check for the end of the sample --
typically just checking the high byte of the zero-page pointer is enough
(in this case, checking $fe). Typically digis are page aligned anyway, so
just zeroing out the unused part (if any) of the last page is fine.

Finally, it is often important that each sample of a digi is played back

at regular intervals. If the samples aren't played at a steady speed,
extra distortion is audible. In the example above, playback is steady
for a full page (256) of samples - but several extra cycles are added
by incrementing the zero-page pointer to the digi. The situation
worsens when we start adding extra code to check for the end of the
digi, and even the main loop starts getting irregular when we add the
code for the simple form of packing discussed earlier (2 4-bit samples
per byte).

These problems can be solved by careful cycle counting and adding NOP
and harmless BIT instructions in strategic places to make each
iteration the same number of cycles, regardless of which branch is
taken - people who have written a stable raster routine, or done some
Atari 2600 coding have likely done this sort of painstaking work
before.

NMI-driven digis

More commonly, however, we enlist the help of CIA #2 and have it
generate regular Non-Maskable Interrupts which we use to call our digi
player. This has two important advantages - first, it makes timing
much more simple. Second, it frees your main program to do other
things while the digi is playing "in the background".

To experiment, I pulled a 4-bit packed digi from the extras disk
included with Super Snapshot 5.22. It's the beginning seconds of the
introduction to Classic Star Trek (Space, the Final Frontier).

Here's the source for a fairly "frills-free" NMI based digi player,
with my comments after blocks of code:

```
start    = $1400
end      = $7cff
freq     = 141
ptr      = $fd
```

Labels start and end simply point to the beginning and end of the
digi. Freq isn't actually the frequency - it's the number of
processor cycles between interrupts necessary to play the digi at the
desired speed/pitch. If you know the frequency (in hz) of the digi,
simply divide your CIA clock speed (approximately 1000000 hz) by the
digi frequency. In this case, the digi runs at approximately 7100 hz.

We use two zero page locations to form a 16-bit pointer to the current
sample in the digi to play.

```
         *= $1000

         ;disable interrupts
         lda #$7f
         sta $dc0d
         sta $dd0d
         lda $dc0d
         lda $dd0d
         sei
```

This code simply disables interrupts and initializes both CIA timers.

```
         ;blank screen
         lda $d011
         and #255-16
         sta $d011
```

Just like erratically timed code can introduce distortion when a digi is
played back, the VIC steals cycles from the processor that can cause
interrupts to not occur precisely when you'd like them to. This routine will
work without the screen blanked, but the extra noise introduced when the
screen is on is noticeable when the time between samples is less than around
2.5-3 times the time the processor is stopped. Another option is to use some
multiple of the raster timing as the sampling rate, and start the routine on a
non-badline, to ensure that the interrupts never occur on a badline. (A final
option is to use a raster-driven interrupt for the digi; with the SCPU, it is
actually possible to drive an IFLI display and play a digi at the same time,
badlines and all -- email Robin for more info, or maybe wait for a future
article!). But the simplest thing to do is to blank the screen :).

```
         ;switch out roms
         lda #$35
         sta 1
```

```
                ;point to our player routine
                lda #<nmi
                sta $fffa
                lda #>nmi
                sta $fffb
```

Unless using the KERNAL routines is necessary in my program, I always
switch out the ROMs. One of the biggest benefits is that our NMI
routine will be immediately called, rather than using $0318/$0319 and
waiting for the KERNAL to indirectly call your routine.

```
                ;initialize player
                lda #<start
                sta ptr
                lda #>start
                sta ptr+1

                ldy #0
                sty flag
                lda (ptr),y
                sta sample
```

This section simply initializes the various memory locations that the
player uses - sets ptr/ptr+1 to point to the beginning of the digi,
loads the first sample, and clears the flag that handles the
alternating between the lower and upper nybble of the packed samples.

```
                ;setup CIA #2
                lda #<freq
                sta $dd04
                lda #>freq
                sta $dd05
```

Sets Timer A on CIA #2 to freq.

```
                lda #%10000001
                sta $dd0d
```

Enables Timer A interrupts on CIA #2.

```
                lda #%00010001
                sta $dd0e
```

Sets Timer A to run in continuous mode. As soon as Timer A counts
down to zero, it will automatically be reloaded to the last writes to
$dd04/$dd05 and begin counting down again.

```
endless  jmp endless
```

For this example, we just put the computer in an endless loop.

```
nmi
                pha
                txa
                pha
                tya
                pha

                ;play 4-bit sample
                lda sample
                and #15
                sta $d418
```

We play the sample while all the code is still linear - before any
branches have occurred. This is to minimize the distorting effects I
mentioned earlier. The AND #15 is used so we don't inadvertently
enable the filter bits in $d418 with the high nybble packed into
sample.

```
                ;clear NMI source
                lda $dd0d
```

By reading $dd0d, we are acknowledging the source of the interrupt,
and the CIA will now generate another interrupt next time Timer A
counts down to zero.

```
                ;just something to look at
                inc $d020

                ;every other NMI do 1) or 2):
```

```
        lda flag
        bne lower
```

Now we deal with "unpacking" the samples.

```
        ;1) shift upper nybble down
upper   lda sample
        lsr a
        lsr a
        lsr a
        lsr a
        sta sample
        jmp exit
```

When flag is set to zero, we shift the high nybble of sample down to
the low nybble so it's ready to be played next NMI.

```
        ;2) get a new packed sample
        ;   then point to next
lower   ldy #0
        lda (ptr),y
        sta sample
        inc ptr
        bne checkend
        inc ptr+1
```

When flag is set to one, we load a new packed sample into sample, and
point ptr at the next packed sample.

```
        ;if end of sample, point to
        ;beginning again
checkend lda ptr
        cmp #<end
        bne exit
        lda ptr+1
        cmp #>end
        bne exit

        lda #<start
        sta ptr
        lda #>start
        sta ptr+1
```

Simply check for the end of the digi, and if we've reached it, loop
back to the beginning of the digi.

```
        ;toggle flag and exit NMI
exit    lda flag
        eor #1
        sta flag

        pla
        tay
        pla
        tax
        pla
        rti

        ;sample's lower nybble holds
        ;the 4-bit sample to played
        ;next NMI - the upper nybble
        ;holds the next nybble to be
        ;played on "odd" NMIs, and is
        ;undefined on "even" NMIs.
sample  .byte 0

        ;flag simply toggles between 0
        ;and 1 - used to decide whether
        ;to play upper or lower nybble
flag    .byte 0
```

Improving D418 Digis

D418 digis tend to generate a lot of noise, because, of course, the 4-bit
sample resolution. Over the years people have come up with numerous tricks to
improve the sound of a d418 digi; here are some that we know of and have tried.

The first, and most obvious, thing to do is to use the low-pass filter, since
a lot of the noise is at higher frequencies. Unfortunately this won't work,

since the filters occur in SID before the volume amplifier -- all the filters
can do is change the DC offset that makes the digi possible. This trick
will work for methods that use SID voices, however (such as Pulse Width
Modulation, discussed in the next section).

Another trick is to "dither" the sound, as discussed in C=Hacking #11. The
idea here is to generate an intermediate "average" value by toggling between
two values. For example, if d418 is set to '8' half of the time, and '9' the
other half, its 'average' value will be 8.5. So this is somewhat like adding
an extra bit of resolution. In principle, you can extend this further: if it
is '8' one-third of the time and '9' for the remaining two-thirds, the average
value will be 8.66. And so on.

Now, we aren't _really_ increasing the sample resolution here, but are instead
increasing the sample playback rate -- we're playing two samples ('8' and '9'
for example) where before we played just one. Don't get too carried away
thinking about "average" voltage levels (after all, there is an average
voltage for the entire digi but that's not what you hear!) -- what's important
is how well the sampled signal represents the original signal. If the
original signal is rising from 8 to 9 during the sample interval, this type
of trick will work well.

Which leads us to another trick: interpolation. This is really a compression
trick, more than a 'resolution' trick. Let's say that one sample value is 5,
and the next value is 9. It might be reasonable to expect an 'intermediate'
value of 7, to play right after the 5. Once again, the idea is to increase
the playback rate to better-represent the original signal. This type of trick
increases the playback rate without increasing the amount of data -- and as
always, your mileage may vary. Many modern soundcards and CD-players use
interpolation.

Another curious trick is to add noise to the signal -- that is, the 4-bit
sample corresponds to the original signal plus noise. Sometimes, by adding
noise to the signal playback the noise can actually cancel! The 'dithering'
trick above can be viewed in this way.

Boosting 8580 Digis

As most people know, there are 'old' SIDs (6581) and 'new' SIDs (8580), and
$d418 digis do not work right on 8580 SIDs, (such as in the 128D, most 128s,
and the 64C) for the reasons discussed earlier -- the 8580 does not have a
residual voltage leading into the amplitude modulator.

The software fix for this is pretty simple: have SID generate a signal, and
hence a voltage, for the volume register to modify. You can actually use
pretty much any waveform to do this, but a pulse is the simplest, since a
pulse wave just toggles between two voltage levels. Moreover, page 463 of
the PRG says, "The TEST bit, when set to a one, resets and locks Oscillator 1
at zero until the TEST bit is cleared. The Noise waveform output of
Oscillator 1 is also reset and the Pulse waveform output is held at a DC
level." So it's not really necessary to worry about the frequency or pulse
width, by using the test bit.

BUT -- it is very important to set the sustain level to $f. The ASDR envelope
generators generate the voltage. A sustain level of 0 gives no improvement.

So, to 'boost' a digi on a later-model SID, you can just turn on a pulse with
the test bit set:

```
        LDA #$FF
        STA $D406
        LDA #$49
        STA $D404
```

Setting more voices gives the digi a substantial extra boost:

```
        LDA #$FF
        STA $D406
        STA $D406+7
        STA $D406+14
        LDA #$49
        STA $D404
        STA $D404+7
        STA $D404+14
```

The moral is: if you're writing a digi routine, and want it to work on all
computers, be sure to boost the digi.

And for completeness, using more channels is a commonly used trick to enhance

digi resolution on the Plus/4. The TED digi resolution (the volume register)
is 3 bits. Fortunately, all channel on/off bits + the volume level are in the
same register ($ff11). If one source is on, the output DC is about half of the
level when both are turned on. This trick can be extended further to results
in a 'semi 4-bit' or 5-bit digi table (the dynamic range is enhanced, but
there are larger steps at the table end than at the start). This trick could
also be used in SID if the sound sources were accurately preset, but runs into
problems due to the non-matching SID-versions and having the control bits in
multiple registers.

SID Type Auto-Detect

The following routine will detect what type of SID is in use. I've
tested it on a fair cross-section of my collection of computers - my
NTSC 128D, two 64Cs, two "breadbox" C-64s, and my PAL breadbox 64. In
all cases the code performed 100% accurately - but still, there may be
cases where it fails. I'd be interested to know if anyone finds any
faults in the routine, so I can improve it!

How does the routine work? I was told that the old SID (6581) and the
new SID (8580) behave differently when set to play combined
waveforms. I coded a fairly simple routine to use the REU to sample
$d41b (the upper 8 bits of Oscillator 3's waveform output) for a full
64k bank. Then I experimented with various frequencies and
combinations of waveforms on Oscillator 3 until I found consistently
different results with the two different SIDs.

When I combined the triangle and sawtooth waveforms and then sampled
$d41b I found that most of the time the oscillator was just putting
out zeros, with occasional bursts of numbers. These "bursts" were
consistently near $ff on the 8580, while the 6581 was always well
below $80 - often $3f was the highest it would get.

So, the detection code ended up being quite simple - I'll explain each
block of code:

        *= $4000

start   sei
        lda #11
        sta $d011

Disable bad-lines (by blanking the screen). This prevents badlines
from interfering with the detection process.

        ;sid setup here!
        lda #$20
        sta $d40e
        sta $d40f

Set Oscillator 3's Frequency Control to $2020. I just randomly chose
this value when experimenting, and it worked, so I kept it. The trick
here is to set a value fast enough that the oscillator will make a
number of cycles (so we can get a good sample of the values coming
out) but not so fast that it might miss any of the "bursts" I was
mentioning earlier.

        lda #%00110001
        sta $d412

Combine the triangle and sawtooth waveforms and start the ADSR cycle.

        ldx #0
        stx high

loop   lda $d41b
        cmp high
        bcc ahead
        sta high
ahead   dex
        bne loop

This loop takes 256 samples of Oscillator 3's output, saving the
highest value in location high.

        lda #%00110000
        sta $d412

Stop Oscillator 3.

```
        cli
        lda #27
        sta $d011
```

Turn the screen back on.

```
        lda high
        rts

high    .byte 0
```

Return from the routine with the highest value sampled from Oscillator
3 in the accumulator. This allows you to branch based on the high
bit:

```
        bmi SID8580
        bpl SID6581
```

Voila!

=====================
Pulse Width Modulation
=====================

        The primary limitation of using the volume register is, of course,
that it is only 4-bits. Pulse width modulation (PWM) allows us to get
around that limitation.

        In general, there are lots of ways of transmitting information.
If you've ever used a radio you've encountered both amplitude modulation,
where the signal is encoded as the amplitude of some carrier wave, and
frequency modulation, where the signal is encoded by changing the frequency
of the carrier wave. In both cases, the idea is to strip out the encoded
information and throw away the carrier.
        Yet another possibility is pulse width modulation: use a pulse
wave at some carrier frequency, and modulate the pulse width. Pulse width
modulation has several nice properties for transmitting signals; we can
take advantage of it to play digis.

        Pulse waves, of course, take on only two possible values: zero and
one (low and high, etc.). Over a single period, a pulse wave will in general
be low for some amount of time and then high for some amount of time.
The _duty cycle_ of a pulse wave is the amount of time it spends in the high
state compared to the total period. For example, a square wave, which is
low exactly half the time and high the other half, has a duty cycle of 50%:

```
             ___     ___     ___
            |   |   |   |   |   |
        ____|   |___|   |___|   |____  ...
```

Remember that, regarding SID, a signal like the above is simply a voltage
level. What is the _average_ voltage over a single period? Since a square
wave is zero half the time and one the other half the average value is
just 1/2. If instead the pulse had a duty cycle of 75%, it would be low
for 1/4 the cycle and high for 3/4, giving an average value of 3/4.

So the _average_ value of a single pulse is simply the duty cycle. So if
we change the duty cycle for each pulse we can essentially generate a
series of average voltage values -- and since a digi is nothing more than
a series of average signal values, we can use PWM to play a digi.

To make this more precise, let's say we had a digi sampled at 1KHz -- one
thousand samples per second. Since each sample value will be approximated
by a pulse, we need one thousand pulses per second. The duty cycle of
the first pulse will be the first sample value, the duty cycle of the
second pulse will be the second sample value, and so on. Note that
the sample rate is the carrier frequency -- the frequency of the modulated
pulse train, 1KHz in this case.

(Actually, to be more accurate, we need _at least_ 1000 pulses per second --
for example, we could use 2000 pulses per second, and represent each sample
value using two pulses. So the more correct statement is that the pulse
carrier frequency is the maximum sample playback frequency.).

The advantage for playing C64 digis is that we have much more resolution
for the pulse width, and probably not in the way you think! Because you
are probably thinking that SID has this nice 12-bit pulse width that

we can use here. The problem is that the absolute highest frequency SID
can produce, using the frequency registers, is about 4KHz, which would
be the maximum playback rate.

There's still another catch -- the carrier wave is still there! Imagine
trying to encode a signal that was constant, say 1/2 everywhere. To
generate a "digi" value of 1/2, you'd use a square wave, half down and
half up. So while the _average_ value of each pulse would be 1/2, the
actual signal would be a square wave at the carrier frequency (look at
the little picture above if you don't see it -- its average value is 1/2).

Trying to modulate a 4KHz carrier wave results in a piercing 4KHz tone,
and a _maximum_ sample rate of 4KHz (and this assumes that you can sync your
code up exactly with SID). So that's pretty worthless for digis.

                        - BUT -

What if we could change the voltage level manually? Let's say some
hypothetical machine language program toggled the voltage level
on each machine cycle -- the result would be a square wave of
frequency 0.5 _mega_ hertz. Okay, let's say it changed the voltage
level every 10 machine cycles -- the result would be a carrier
frequency of around 50 KHz. The point here is that a machine language
program can generate its own pulse waveform, and do so at much higher
frequencies than SID can produce.

Toggling the voltage levels turns out to be very simple. As was
described earlier, the way to "boost" digis on later SIDs is to use
a pulse waveform at frequency zero. Depending on the value of the
pulse width register, SID will set the output voltage to either high
or low. So all a program has to do is set up a pulse waveform at zero
frequency and use the pulse width registers to toggle the voltage --
set $d403 to either $00 or $ff to toggle low/high. (You could also use
$d418 to toggle low/hi, but this method should produce more uniform
results, and unlike $d418 can be filtered).

So now we're cooking -- we've got a program that can generate a pulse
train. The next step is to change the width of each pulse to represent
the sample values in our digi. Remember that the duty cycle -- the
percentage of time the pulse spends high -- is the average value for that
pulse. But also remember that each digi sample represets an average
value over the sample period. If the pulse period is equal to the sample
period, then _the duty cycle is exactly the sample value_!

Example: let's say that we have an 8-bit sampled digi, so that values go
from 0-255, and our program generates pulses with a period of 256 "ticks".
Now pick a sample value, say 56. All the program has to do is hold the
pulse high for 56 "ticks", and low for the remaining 255-56 = 199 "ticks",
and it will have the correct average value: 56/256. So a program to play
8-bit samples might look like

1 - Load .X with next sample value
2 - Load .Y with 256-.X
3 - Set pulse high
4 - Loop for .X iterations (each loop iteration is one "tick")
5 - Set pulse low
6 - Loop for .Y iterations
7 - Loop back to step 1

Let's say that each "tick" takes m cycles, and the sample size is 2^n, so
that there are 2^n ticks per sample. A stock machine runs at around
10^6 cycles/second, so...

        (10^6 cycles/second) / (2^n ticks/sample * m cycles/tick)
        = 10^6 cycles/second / (m * 2^n cycles/sample)
        = 10^6 / (m * 2^n) samples/second

So, for example, let's say we had n=6-bit samples -- 2^6 = 64 -- and could
generate pulses with a resolution of one machine cycle -- m=1. Then
we could play that 6-bit sample at 10^6/64 = 15.6KHz. That is _really
very good_! In principle -- possibly using the CIA timers, possibly using
fixed delay loops, possibly using a massively unrolled loop -- this can
be done on a stock machine. (I did try using the CIA timers, but the
number of cycles to set up the timers was too big, and made it sound poor;
I've included the code below though.)

At this point it becomes a numbers game. As we increase the sample size
(increase m or n above), we _decrease_ the sampling rate -- if, in the
above example, we instead use 8-bit samples, the sampling frequency drops
by a factor of four to around 4 KHz. So there's a tradeoff between
resolution and sampling frequency.

AND... we still have this issue of the carrier frequency. You should be
able to convince yourself that the sampling frequency above is exactly
the carrier frequency. So with the 8-bit resolution example there
would be an awful 4KHz tone running through the playback. There are
only two ways to beat the carrier frequency: push it high enough that
you no longer hear it, or else push it high enough that you can use the
filters to dampen it down.

How high is high enough? You can judge for yourself, but 15 KHz is
pretty tough to hear, unless you have good ears and the volume is really
loud -- so 6-bit samples are within reach on a stock machine.

But add a SuperCPU into the picture, and the numbers get _really_ nice.
Everyone knows that a SCPU can interact with the C64 at 1MHz, and
hence generate pulses with 1MHz resolution, using code like

```
        lda #$ff
        sta $d403          ;Set level high
:loop   lda $d011          ;wait for C64 cycle
        dex
        bne :loop
```

where .X contains the sample value. But what happens if we try to move
beyond that 1MHz? What if we put some NOPs into the above delay loop,
in place of the lda $d011? Well, in principle it means that the duty
cycles won't always be right, which corresponds to some sampling error.
In practice, however, it works _really well_! Consider what happens when
the above code is changed to:

```
:loop
        nop
        nop
        dex
        bne :loop
```

The earlier formula still applies, but now using 20MHz cycles:

$$20 * 10^6 / (m * 2^n) \text{ samples/second}$$

In this example each loop iteration -- each "tick" -- is nine 20MHz cycles,
giving a playback rate of approximately 17Khz for 7-bit samples. Which
is TOTALLY COOL!

And it can even be pushed to 8-bit samples (although I personally don't think
they sound any better, at least with the code I've tried; maybe the code can
be improved). Using loops like

```
:loop
        dex
        beq :done
        dex
        beq :done
        ...
        dex
        bne :loop
:done
```

it is possible to "fine-tune" the loop tick to somewhere between 4-5 cycles,
giving a playback rate between 15KHz and 19KHz, for an 8-bit sample. Pretty
cool. The code is also a little more involved (with 7-bit samples we can
use BMI for the loop branches; not so with 8-bits). But it really is
possible to play 8-bit samples at 19KHz on a C64 (plus SuperCPU).

Using two voices

You may be thinking, Hey, we've got three pulse waves to work with, can
we improve the performance by using multiple pulses?

Let's say we have two pulses, P1 and P2, with the same period. When both
are activated, the pulses simply add together -- that is, the total voltage
is just the sum of the individual voltages, and therefore the _average_
voltage is the sum of the individual pulse averages:

$$\text{avg voltage} = D1 + D2$$

where D1 and D2 are the duty cycles of pulses P1 and P2. In the simplest
case, this gives us an extra bit of resolution -- if D1 and D2 are both
7-bit values, say, then D1+D2 is an 8-bit value.

-BUT-

Consider, for a moment, what would happen if we were to change the amplitude
of the second pulse -- that is, let's say the maximum voltage it took on
was 1/16 of the maximum voltage of the first pulse. The average voltage
would then be

        avg = D1 + D2/16

This then gives us _four_ extra bits of resolution, with each bit to the
right of the decimal place. For example, if D1 and D2 are 4-bit numbers,
with D1=xxxx and D2=yyyy, then the avg will be a number like xxxx.yyyy
(four bits to the left of the decimal place and four to the right).

Of course, we can change the pulse amplitude by changing the sustain
setting, so in principle this gives a very easy and efficient way of
playing high-resolution digis. In practice, I have not been able to
make it work very well. I used a sustain setting of 1 and split an
8-bit sample into two 4-bit pulses; I believe the result sounds better
than 4-bits, but certainly doesn't sound anywhere near 8-bits. My
suspicion is that it is because the second pulse voltage is not really
1/16 of the first pulse, which corresponds once again to adding noise
to the sample value.

To find out, we can just measure the output at different sustain levels.
The following table gives the voltage output for voice 1 using a pulse
waveform at zero frequency and volume 15:

	Pulse Width		Diff	
SU	000	fff	000	fff
0f	6.34	5.29	.08	.07
0e	6.26	5.36	.02	.01
0d	6.24	5.37	.06	.05
0c	6.18	5.42	.03	.02
0b	6.15	5.44	.05	.03
0a	6.10	5.47	.03	.02
09	6.07	5.49	.04	.02
08	6.03	5.51	.03	.02
07	6.00	5.53	.05	.03
06	5.95	5.56	.03	.02
05	5.92	5.58	.05	.02
04	5.87	5.60	.04	.03
03	5.83	5.63	.04	.02
02	5.79	5.65	.06	.02
01	5.75	5.67	.05	.02
00	5.70	5.69		

Voice 2 is identical within a few hundredths of a volt. If this test is
repeated using voices 1 and 2 simultaneously, the result is:

	Pulse Width	
SU	000	fff
0f	7.30	5.25
0e	7.12	5.36
0d	7.09	5.37 (!)
0c	6.95	5.46
0b	6.88	5.49
0a	6.78	5.54
09	6.72	5.58
08	6.62	5.62
07	6.58	5.65
06	6.47	5.70
05	6.40	5.73
04	6.31	5.78
03	6.22	5.82
02	6.13	5.87
01	6.07	5.90
00	5.97	5.95

Note the weird step at $0d -- the response is definitely not linear!

Now, to summarize, when using one voice, the "positive" amplitude (about the
mean 5.70V) is .64V and the "negative" amplitude is .41V, giving a spread of
1.05V. With two voices together, the amplitudes are 1.33V, 0.72V, and 2.05V
respectively. If the two signals were simply added together, the numbers
should be 1.28V, 0.82V, and 2.1V.

What we originally wanted was a signal like

        D1 + D2/16

that is, another pulse that is 1/16 the value of the 'full' pulse. 1/16 of
the positive amplitude is .64V/16 = .04V, and 1/16 of the negative amplitude
is .41V/16 = .026V. A setting of sustain level 1, on the other hand, gives
voltage offsets of 0.05 and 0.02, giving approximately

```
        .64V / .05V = D1 / 12.8
        .41V / .02V = D1 / 20.5
```

So, in summary, whereas I wanted D1 + D2/16, I was actually getting something
that varied from D2/12.8 to D2/20.5, even if the two voices summed together
correctly.

There may still be a way to make all this work right, which would be great,
but I'm tired :). The code from my attempts is below.

I also could not get two 7-bit pulses to sound like an 8-bit pulse. I took
an 8-bit pulse and divided it in half, assiging each half to a pulse
(and giving the extra bit to pulse 2, if an extra bit was present).
I suspect that another issue is that it is impossible to update both
pulses simultaneously, meaning some delay between pulses, which translates
to adding -- surprise! -- noise to the signal. Perhaps it would be
more effective at lower resolutions, however.

If someone has some success using these techniques I'd be interested in
hearing it.

SID lockups

Blindly applying these PWM algorithms has a way of locking up SID -- like,
locking him up hard. To be honest, I don't have a good explanation for why
this happens, and I haven't yet found a good method of prevention -- toggling
the test bit, playing a real sound for a short time, toggling the gate bit,
and so on, just don't seem to "initialize" SID reliably enough. Sometimes the
code works, and sometimes it doesn't -- it's the same code both times. Often
resetting the machine will make things work; I'm not sure what hardware resets
take place within SID, but the kernal certainly zeros him out so that's a
possibility. The other observation is that playing a tune seems to 'clear
out' whatever is blocking SID. So there _must_ be some kind of software
solution to the problem.

In the example code pressing RESTORE restarts the code, which will usually
clear the 'blockage' after a tap or two, if it happens.

If anyone has some thoughts on this issue (or even better, an explanation
of what is going on!) I'd love to hear them.

========
The Code
========

And that about sums stuff up, we think. All that remains is a zipfile
containing code examples of the things we've discussed. The archive
contents are:

```
digi-$1200.o  pwm-2cia.d.s  pwm-jmj      pwm8.d.s
digi-$1200.s  pwm-cia.f.s   pwm.d.s
pwm-$1200.s   pwm-cia.g.s   pwm2.c.s
```

The main example program is "pwm-jmj", for the SuperCPU. This file contains
code at $1000 and a sample at $1200 (the sample is a sample I made on my
Amiga years and years ago, of Jean Michel-Jarre; it's really not very clean,
but it still sounds pretty nifty). So, to run it just load and SYS 4096.

You _need_ a SCPU to run it; it plays the sample at 7-bits and around 16KHz,
using pulse width modulation (pwm-$1200.s is the source code). If you want
to compare it to a 4-bit $d418 digi, just load "digi-$1200.o",8,1 and
sys 4096. The other files are different code experiments -- using two
pulses, using the CIAs, 8-bit samples, and so on (all written in Sirius,
of course) and may provide ideas for anyone wanting to experiment further.

If you want to try your own sample with the code, I suggest converting it
to RAW format. RAW format is just the digi, with no headers, and uses
signed 8-bit numbers. It's what I converted the sample to, and what the
code is designed to play!

Okay, enough talk -- go listen to that cool 7-bit 16KHz JMJ digi!

```
begin 644 digibin.zip
M4$L#! H    !    -IQQIT41>R90   &L      ,    9&EG:2TD,3(P,"YD(P""D
```

```
MYW5P/#3.N/_[W. $,#_04M15O>#AX@8SL/A!"::6KXSG#E2HAW:DP,;?,ZH+
M$S589G187]-WX\8#1WQPN5A*15V[YC]/>WI,I!]*_@3H$3- F3'LFOH5$A!L
M $H0^!_TF74%S(K37^Q9TO^Z;03%,M^SGN!@E(= 0P[]M4D#3TASXKN4T9[^
MONEBA(W99QBELG0%-17@:8M!)EV;DP&O*_^6( NNXX!69)T^>WF KU!+ P0*
M 0 !@#Z:,<:33MY=.T! !4 P !P '!W;2YD+G-*!QB= _T= QT(!4T&
M""T'% 4(!@@'%@@&&!2@=!"!#4!!18#%&0%) @$ P0%!@@&!1@-%_V=_OYN!/[^
M+A<>!PXG!1X&!'8!@!!'@=^!?[N[$N#$P!P(# @@#B!C>C+-,!QB%!'#2?!O[^
MCIO_TK^^^#G7==C=@N]!<3.,MyQ... (garbled)
```

```
end

.......
....
..
.                                       C=H #20

:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::

cu next time!

-S
.......
....
..
.                                       - fin -
```