Linux/Alpha活用講座

Linux/Alphaによる数値計算ならびにRISCアーキテクチャのおはなし

第10回　Alpha21264とIprobeのオプション

清水尚彦 nshimizu@et.u-tokai.ac.jp

はじめに

7月号でCompaq社からコンパイラが出るという噂があると書きましたが、私自身、複数のルートからそういった話を聞きましたので、ほぼ確実な情報のようです。ただ、人によってFORTRANが出ると言っていたりCが出るといったりとまちまちなので、本当の所は良く分かりません(両方でるのかな？)。いずれにせよ、CompaqがますますますますLinuxへのコミットを強めるようで心強いですね。LinuxとTrue64 UNIXの差は、数値計算に関する限り、コンパイラとライブラリの違いが一番のポイントだったのですから、この2つがクリアされれば、数値計算にLinuxを使わないという理由はなくなります。もちろん、サードパーティのベンダーソフトを使いたいという向きにはLinuxはちょっと困るので、True64自体が不要となっているわけでは全然ありません。

最近ではすっかり以前の勢いが無くなりましたが、私はNOS(Network OS)の世界では、NovelのNetWareが好きで、業務でシステムを構成するならNetWareに限ると思っていました。ユーザーが使う端末は台数が多いので、その保守費用はばかにならないものがあります。「機械物」の常として、可動部分が多くなると故障が増えるのは常識なので、ハードディスクもフロッピーディスクも載せたくないので、ディスクレスの構成が取れるOSが必要だったからです。さらに最近では、ソフトウェアライセンスの問題があり、困ったことにWindowsを搭載したマシンの台数が増えると、勝手にソフトウェアをインストールするようなユーザーの行動を管理しきれなくなります。

1つの方法として、LinuxかFreeBSDをクライアントとして、ブートROMをいれたカードからディスクレスブートさせて端末とする方法があり、NetWareに代わるNOSとして、業務システムでまともに使える唯一のエンドユーザーOSとなる可能性があると思っています。これは業務システムだけでなく数値計算のシステムでも同様のことが言えます。高信頼性のシステムを構築するには、なるべく稼働部分のあるリソースは減らすべきなのです。Sandia National Laboratoriesの「CPLANT」は、まさにそのようにしており、計算ノードである400台は、ディスクレスでLinuxをネットワークブートしています。CPLANTの話はまた稿を改めて紹介したいと思います。

ところが、アプリケーションの面から考えると困ったことに、Windowsの導入が避けられない場面が少なくありません。デュアルブートは前に示した目的に沿っていないことは明らかだと思います。そこで、他の方法を模索しているわけですが、以前、「WinCenterPRO」という、WinFrameに皮をかぶせたシステムを使い、その快適性に満足しておりました。しかし、これには日本語が使えるバージョンがなく、導入目的には合わなくなります。ずっと前に、この日本語版が出ないのかと問い合わせてみたことがあります。そのときは、なかなか良い返事がもらえなかったのですが、実は開発元のCITRIXがマイクロソフトに技術供与し、マイクロソフトから「Windows NT Termianl Server Edition(TSE)」として出荷されたのでした。

このTSEですが、CITRIXとの契約がどうなっているのか分かりませんが、CITRIXお得意のマルチプラットホームのプロトコルを使わずに、別のプロトコルでクライアントと通信することになっているのです。そのため、TSEだけで接続できるクライアントはWindows 95／98とWindows NT、そしてこのためにマイクロソフトが発表したWindows Based Terminal（WBT）だけになってしまいました。マルチプラットホームにするにはCITRIXから「MetaFrame」というソフトを買わなくてはならないし、WinCenterPROで提供されていたNISやNFSの機能は「WinCenter for TSE」というソフトになって、やはり別途購入しなくてはならなくなっています。

WBTをユーザーに持たせて、Windowsはサーバの1台のみにして、かつ、各クライアントマシンではCD-ROMもフロッピーディスクも使わせないことにし、モニタ無しのサーバにてプログラムを実行することにすると、Windowsアプリケーションに限っては、このWindows NT TSEの運用形態が、私のイメージにぴったりなのです。が、この雑誌の読者の皆さんならよくご存知のように、Windowsだけで満足できる環境にするのは大変難しのです。このように、相変わらずWindows環境とLinux環境の混在にこだわっている私は、Windows NT TSEとMetaFrameとWincenter for TSEの見積りを取ってみました。しかし、結論から言うと、あまりにソフトウェアが高額で、とてもじゃないけど導入の検討対象にならないです。

MetaFrameで一番安価な5ユーザーのものを前提に考えてもアカデミック割引をしても1ユーザーあたり20万円程度の負担になってしまいます。TSEだけなら我慢できる範囲の金額に収まりますので、これらの付加ソフトの価値がそれに見合っているのかどうかが問われます。ユーザー管理を別々にすることにしてWincenterをあきらめてもまだまだ高価ですが、MetaFrameを止めると、Windows NTかWindows 98、もしくはWindows Based Terminal（WBT）以外からは接続できないことになってしまいます。これらのソフトの関係は次のようになっています。

TSE マルチユーザー用のWindowsNTサーバ
MetaFrame マルチプラットホーム向けプロトコル変換サーバ
Wincenter NISやNFSによるUNIX環境との親和性向上

TSEはないと始まらないのでしかたないのですが、これだけでは、プロトコルとしてMicrosoft独自の「Remote Desktop Protocol（RDP）」しか使えず、Linuxからの利用はできません。RDPはT.120プロトコルをMicrosoftで改良したものということですが、仕様がオープンでないので、Linuxに実装される可能性は低いです。とすると、TSE＋MetaFrameが最低必要になってしまい、高価なMetaFrameが避けられないということになります。

ですが、ここで発想を少し変えて、クライアントからXとNTが使えれば良いのだとすると、TSEの上にXサーバを載せれば十分と言うことになります。すると高価なMetaFrameの変わりに、XサーバとTSEを購入して、クライアントにはWBTを導入すればよいということになります。しかし現在では、WBTは極めて高価で、同じ性能ならPCを買ったほうがずっと安いということになっていますので、TSEで接続させるためにWBTを導入するのはペイしない計画になってしまいます。それでも、維持費等を考えれば、TSEのメリットが活きてきます。このような模索をしている中で、高岳製作所の「MiNT ACC」という端末は、WBTとしてもX端末としても使えるという触れ込みがなされていたので、さっそく資料を請求しました。しかし、これらの機能はブート時に切り替えるようになっていて、混在はできないようです。これでは私の要求は満たせないので別の手立てを考える必要があります。

このように、何とかしてTSEの便利さとLinux環境への安価な適用を果たしたいといろいろ考えていますが、実現性を無視した「wishing list」としては

- LinuxもしくはFreeBSDベースのRDPクライアントソフトウェア
- ディスクレスブータブルなWindows NT or Windows 98
- X＋WBTの同時利用可能な端末
- Wine上でRDPクライアントを動かす（これが現実的？）

などがあるといいなぁと思います。このほかに、最近流行の「VMware」を使うと、次のようなことができるかもしれないと考えています。

- Linuxをディスクレスブート
- NFSにインストールしたVMwareからWindows 98をブート
- Windows 98のRDPクライアントソフトでTSEに接続

このやり方なら、ディスクレスなPCからTSEを利用できるので、WBTのようなものを買わなくても済みますし、何より、Linuxを同時に利用できるので、大変優れた方法だと思います。VMwareがあまり多くのディスプレイ

ドライバをサポートしていないので、画面の解像度には不満が出るかもしれませんが、どうせ私にとってWindowsアプリケーションなんて、PowerPointとブラウザくらいしか役に立たないのだから、解像度は割り切って考えることもできます(ただし、ユーザに勝手にインストールさせないという目的は達成困難になる可能性が大きいです)。

VMwareだと、何でこんなことができるのかというと、VM(Virtual Machine：仮想機械)を提供することで、VMの上にWindowsをインストールできるからです。DOSEmuと異なり、エミュレータではないので、特権命令以外の命令のオーバヘッドは比較的少なくて済みますから、性能も悪くない可能性が高いです。VMwareは使ったことがないのですが、仮想計算機の考え方はどこでも似たようなものです。

私は昔、メインフレームの仮想計算機構の開発を担当たこともあり、仮想計算機にはちょっとうるさいのですが、Alphaの世界にはVMwareはないので、フリーなVMCPを作ってみたいと思います。メインフレームと異なり、元々、仮想計算機を前提には設計されていないので、仮想計算機を作るのに必要なハードウェアの機構がAlphaのアーキテクチャに備わっているかどうかは慎重に検討する必要があります。ですが、自己仮想化が難しいと言われていたx86でもできているのだから、Alphaでできない理由は少ないと思います。

もちろん、難しい機械で無理に仮想計算機の導入を計ろうとすると、本来高速であるべきユーザーモードのプログラムも遅くなるような実装となる可能性は高いです。それに、x86ではハードの使い方が軽いWindows95／98を対象にできるのに対して、Alphaでは、Windows NTを対象にしなくてはならないので、仮想計算機への要求仕様が若干高い可能性があります。しかし、ARCのPALコードをエミュレートして画面をVGAとすれば、Windows NTを仮想計算機上で動作させることもできる気がしています。一度にこのあたりの確認をするのは大変なので、連載の中で仮想化の可能性を少しずつ探っていくことにしたいと思います。

最近のAlphaの話題から

●Alpha 21264

さて、先日Webを見ていたら、21264のデータシートが出ていました。ハードウェアリファレンスマニュアルとは異なり、データシートは物理仕様を中心に記述されているのですが、一部は内部構成の話も出ていました。おそらく内部構成の話をデータシートに書いたらハードウェアリファレンスマニュアルは個別には出ないと言うことなのでしょうね。と、ここまで書いているときに締め切りは過ぎてますが(m_o_m)別の資料が出てきました。これはコンパイラ製作者へのガイドというもので、データシートと同じホームページからダウンロードできます。

ざっと眺めると、こちらには21264の内部アーキテクチャが詳細に書かれています。マザーボードの設計者以外はデータシートでなくこちらのガイドのほうが役に立つと思います。この資料に付いて詳しく調べていると、また連載がスリップする可能性があるので、後のお楽しみということにさせてください。これらの資料に興味のある方は次のURLをご参照ください。

```
http://ftp.digital.com/pub/Digital/info/semi-
conductor/literature/dsc-library.html
```

さて、21264はアウトオブオーダ実行の機構を採用したおかげで、命令の並びは直接的には性能低下要因にはなりにくくなりました。これはコンパイラの最適化がいまいちでも、ハードウェアが勝手に命令の実行順序を入れ換えられるからです。

そこで、21264を搭載したマシンは商用のコンパイラを使わなくても比較的性能的を出しやすく、一般的な用途には他のプロセッサより優れているので、私としてはお勧めします。値段もずいぶんこなれてきました。執筆時点ではDCGという会社が、$3,500で販売しているそうです。URLは、

```
http://www.dcginc.com
```

です。

お勧めのマシンは、433MHzの21264を搭載した「DS10」というマシンです。ただし、これはLSIの数を減らしてデータバスの転送スループットを下げているので、それなりの性能ペナルティがあります。つまりクロック周波数の差以上に性能が下がっています。それでもSPECfp95は40台の後半にあり、同世代の他のアーキテクチャより圧倒的に高い数値になっています。一番安価なシステムではグラフィックスカードも入っていない構成ですが、サーバ用途に使うには十分でしょう。Linuxの以前の

カーネルだと、ディスプレイレスでインストールするのは難しいのですが、カーネルソースを展開すると“Documentation”のディレクトリに、“serialconsole.txt”というドキュメントがあって、ディスプレイレスのシステムも比較的容易に構築できるようになってきたようです。AlphaのSRMコンソールのように、シリアル端子から制御できるようになっているBIOSを搭載していれば、サーバ目的にはディスプレイなんてないほうがずっと便利です。

●スーパーコンピュータ・ランキング

さて、毎年2回、世界のスーパーコンピュータサイトの計算能力を上から500位までリストアップするお祭りが行われます。昨年末のリストには、Linux/Alphaを使ったシステムが2つ入っていましたが、6月に発表された新しいリストでも、これらのサイトはリストに残っていました。ただし、順位は昨年より下がっていて、この世界の競争が激しいことを物語っています。

この計算能力は一般的なものを出すのではなく、単に「LINPACK」と呼ばれる密行列の係数行列を有する連立一次方程式の解を求める時間の速い順に並べただけのものです。ですから、ここで上位に位置するからといって必ずしも高性能なシステムであるとは限りません。

それではこのリストを少し眺めてみましょう。まず、Linux/Alphaを使ったシステムの内容は表1の通りです。

このランキングを、Linux/Alphaから離れて、Alphaのシステムとして見ると興味深いものがあります。

表2には、Alphaを使ったシステムの数を示しました。このように、大体どの順位の範囲を取ってみても、その約半分はAlphaを用いたシステムになっています。しかもこの範囲のシステムはすべてSGI/Crayの「T3E」とその系列のシステムになっているのです。ということは、TOP 500に載せる一番簡単な方法は、SGI/CrayのT3Eシリーズを購入することだと言えましょう。

●Iprobeのオプション

前回の7月号で「次回はIprobeの話をする」と書きました。ちなみにIprobeは、Alphaの性能を測定するためのパッケージで、Compaq自らが公開しているツールです。これにより、様々な角度からAlphaプロセッサの性能を測定することが可能となります。

一応動作させてそれなりの性能指標が得られるようにはなったのですが、思ったところにぴったりはまらない部分が多く、記事としてまとめるのはもう少しお待ちください。7月号で示したオプションの類は、実は実際のコマンドではドキュメントと少し異なっていて、びっくりする方もいるかと思いますので、「iprobe -help」として得られるデータをリスト1に示します。

見て分かるように7月号の説明と違い、パラメータのスペルが省略形になっていたりするのでご注意ください。

私はビジネスマシンの設計が長いので、ついつい視点がシステムやビジネスの面に行くことが多いです。純粋数値計算の人達から見ると邪道で蛇足的な話が多いかも知れませんが、ビジネス分野の利益率がHPC分野の開発を現実には支えているともいえる状況があるので、ビジネスの視点は大事だと思っています。ビジネスなんて関係ないやといっていては金の卵を生む鶏を殺すことになりかねません。Linux/Alphaでも、ビジネス分野の応用も広げて行きたいと思います。

表1

ランキング入りしたマシン	説明
Sandia National LaboratoriesのCPlant Cluster	これはマシンとしては、計算ノードとして400台の21164Aを用いていますが、ベンチマークには、150台のシステムの報告をしているだけです。計算ノードは、本文にも書いたように、ディスクレスでMyrinetを用いてスイッチ接続しています。このプロジェクトは計算センターのような運用を目指しています。 昨年のリストと同じ台数で同じ性能を報告しているのですが、これは彼らがこのお祭り騒ぎに新しい結果を報告していないだけでしょう。順位としては129位となっていて性能は54.24GFLOPSとしています。
Los Alamos National LaboratoryのAvalon Cluster	140台の21164AのLinux/Alphaを載せたPCを、100BASE-Tのスイッチで接続しています。CPlantと異なり、安いハードからなるべく性能を引き出そうというシステムになっています。順位は160位で性能は48.6GFLOPSとなっています。 CPlantとの差はわずかですが、順位は大きく下がっていてこのクラスの競争の厳しさを示しています。実はこの間にあるマシンの大半はSGIの「Origin 2000（250MHz）×128ノード」のクラスタで、こちらが51.44GFLOPSとなっているのです。

表2

順位	システム数
1 - 10	5／10
1 - 20	11／20
1 - 40	22／40
10 - 20	7／10
21 - 40	11／20

リスト1　Iprobeのオプション(「iprobe -help」の実行結果)

```
IPROBE V4.00 compiled on Jan  6 1999 at 09:24:46
      Digital Equipment Corporation,
      Copyright (c) 1993, 1994, 1997, 1998

Usage: iprobe [-flag [-flag...]] [eventspec [eventspec...]]
 flags, one of:

        -buffer_count
                    number of buffers to allocate for PC
samples (sampling only)
        -delay_start
                  # of seconds between tool start and start
of measurements
        -duration  length of time, in seconds, for tool to
measure
        -help      Display event, help, switch and/or usage
help
       -interval  number of seconds between reports (counting
only)
        -method    type of measurement to perform
        -command   command line for starting a process to
profile
        -mode      record events in the specified mode
        -output    name of output file in which to place
information
        -pc_range  add an addr range and granularity to the
addr histogram list
        -quiet     Do not print machine/argument summary
prior to measurement
        -no_intermediate
                  Do not print intermediate results during
measurement (applies only to counting at present)
        -buffer_size
                   Size of sample buffers to be allocated
        -input     name of file to receive command line
arguments from
        -collect   Collect the following data in sampling
        -nocollect Do not collect the following data in
sampling
        -add       Collect the following data in addition to
default data



        Flag name                              Default
        ~~~~~~~~~                              ~~~~~~~
        -buffer_count integer                  3
        -delay_start integer                   0
        -duration integer                      0
        -help                                  (null)
        -interval integer                      1
        -method measuring-type (R)             count
        -command quoted string
        -mode mode-type (R)                    all
        -output filename                       pcsample.dat
        -pc_range range-spec (R)               0-;8192
        -quiet
        -no_intermediate
        -buffer_size integer                   page length
        -input filename                        iprobe.cmd
        -collect filename
        -nocollect filename
        -add filename

      flags should begin with either '-'

      flags marked (R) can be specified multiple times

      eventspec:
        event[.freq][.skip]
```

↘

```
      event is the name of the event to be measures, and
must be one of
        those appearing in the list that follows.

      freq is the frequency at which interrupts are to be
generated
        and can be HIGH, LOW, or a value.

      skip indicates that all but the skip-th interrupt is
to be
        ignored and must be a value. (ignored for all but
sampling)

      The valid measuring-types are:

        sample     Create a file of PC samples based on the
specified events
        count      Periodically print the number of specified
evts/second
         total       Periodically print histogram of the
specified evts/second
        ipl        Periodically print histogram of the # of
evts at each ipl level
        mode       Periodically print histogram of the # of
events in each mode
        address    Print a histogram of the specified address
ranges
         priority   Print a histogram of system based on
priority

      Finally, the valid mode-types are:

        kernel       Measure events in kernel mode
        user         Measure events in user mode
        pal          Measure events in PAL mode

      The various data types that can be collected are as
follows.
      These options can be used with following options
      -collect -add and -nocollect.

        ps           Collect PS low longword
        pc           Collect PC low longword
        time         Collect the sampling time
        pid          Collect current process id
        ctr          Collect counter number
        pri          Collect current process priority


Events defined on the current system -- select up to one
event from each column:

   *issues                          dual_issue_cycles
   *non_issues                      branch_mispr
    pipe_dry                        integer_ops
    pipe_frozen                     float_ops
    branches                        stores
    loads                           icache_miss
    cycles                          dcache_miss
    palcode_cycles                  bcache_miss
    bcache_victim

    Frequencies                     Frequencies
    Low :65536                      Low :4096
    High:4096                       High:256


      An asterisk indicates the event is measured at twice
the specified frequency.
```

↗