Linux クラスターによる教育・研究両用システムの構築

CAUA 第一回研究教育分科会 (2000 年 10 月 13 日) 予稿

宮澤 三造 (群馬大学工学部共通研究室, miyazawa@smlab.sci.gunma-u.ac.jp)

要　　旨

工学部学生の情報処理演習用に使用する 120 台の Linux PC を、演習用のデスクトップとしてのみならずクラスター構成により研究用に使用するシステムを構築運用している。クラスターとして使用するためには、24 時間稼働は前提である。リブートできないようなハード、ソフト両面での設定も必要である。システムの概要をご報告する。

目次

1 X 端末と PC Unix/Linux の比較

	X 端末		スタンドアローン
ディスク	不必要		必須
音声	不可		可能
アプリケーションサーバーへの負荷		>	
応答		<	
アップグレード (X サーバー /Window Manager)	困難		可能
機器の有効利用		<	
ハード性能		<	
価格		>	
納入時管理コスト		∼	
維持管理コスト (アップグレード無しの場合)		∼	

2　ハードウエア構成、納入時のハードウエアチェック及びセキュリティー対策

2.1　ハードウエア構成

	PC x 120	PC server x 6	Alpha x 18
cpu	Celeron 500MHz	Pentiume II 750 MHz	Alpha 21264A 750MHz x 2
mainboard	A Open AX6BC Type R (Intel 440BX)	TYAN Tiger 100	DP264
memory	ECC 128 MB	ECC 1 GB	ECC 1 GB x 16/2GB x 2
system disk	Seagate (ST36421A) ATA 6 GB	Ultra 160 9 GB	Ultra 2 Wide 9GB
network	3Com 3C905B (10/100Base TX)	3Com 3C905B	3Com 3C905B
video card	Matrox G400	Matrox G200	3Dlabs
sound card	約 20 台程に CREATIVE Vibra 128	-	-

2.2 納入時のハードウエアチェック

納入前のハードウエアチェック

- メモリーの read/write テスト
- ディスクの read/write テスト

2.2 納入時のハードウエアチェック

納入前のハードウエアチェック

- メモリーの read/write テスト
- ディスクの read/write テスト

納入後のハードウエアチェック

- リブートテスト

2.2 納入時のハードウエアチェック

納入前のハードウエアチェック

- メモリーの read/write テスト
- ディスクの read/write テスト

納入後のハードウエアチェック

- リブートテスト

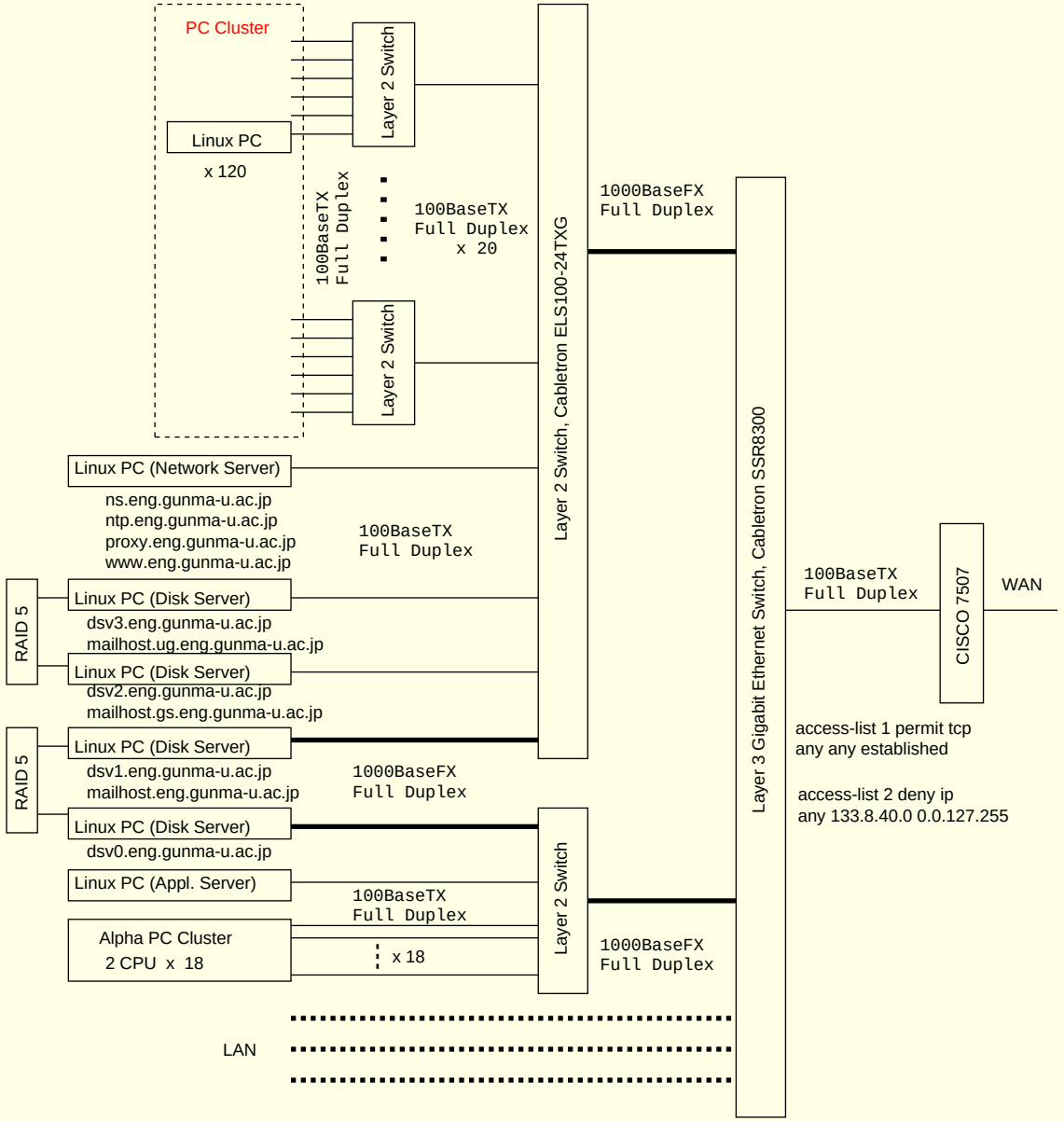
メインテナンス

- 予備機を用意し send-back 方式

2.3 デスクトップのセキュリティー対策

ハードウエア	
CD-ROM	無し
Floppy disk	無し
筐体	特殊ネジにより固定
リセットボタン	無効に設定
電源ボタン	無効に設定
ソフトウエア	
BIOS	パスワードによりロック
	キーボード無しでもブート可能に設定
Boot loader	GRUB
	パスワードにより Multi-user mode 以外ではブート不可能にロック
CTRL-ALT-DEL	inittab で無効に設定

3　ネットワーク接続


PC Cluster
Linux PC
x 120
Layer 2 Switch
100BaseTX
Full Duplex
100BaseTX
Full Duplex
x 20
Layer 2 Switch
Layer 2 Switch, Cabletron ELS100-24TXG
1000BaseFX
Full Duplex
Layer 3 Gigabit Ethernet Switch, Cabletron SSR8300
100BaseTX
Full Duplex
CISCO 7507
WAN
access-list 1 permit tcp
any any established
access-list 2 deny ip
any 133.8.40.0 0.0.127.255
Linux PC (Network Server)
ns.eng.gunma-u.ac.jp
ntp.eng.gunma-u.ac.jp
proxy.eng.gunma-u.ac.jp
www.eng.gunma-u.ac.jp
100BaseTX
Full Duplex
RAID 5
Linux PC (Disk Server)
dsv3.eng.gunma-u.ac.jp
mailhost.ug.eng.gunma-u.ac.jp
Linux PC (Disk Server)
dsv2.eng.gunma-u.ac.jp
mailhost.gs.eng.gunma-u.ac.jp
RAID 5
Linux PC (Disk Server)
dsv1.eng.gunma-u.ac.jp
mailhost.eng.gunma-u.ac.jp
1000BaseFX
Full Duplex
Linux PC (Disk Server)
dsv0.eng.gunma-u.ac.jp
Linux PC (Appl. Server)
100BaseTX
Full Duplex
Alpha PC Cluster
2 CPU x 18
x 18
Layer 2 Switch
1000BaseFX
Full Duplex
LAN

4　システム設定

4.1　システム設定の方法

一台の PC をシステム設定した後、そのディスクのデッドコピーを納入依頼。

4 システム設定

4.1 システム設定の方法

一台の PC をシステム設定した後、そのディスクのデッドコピーを納入依頼。

4.2 必須なシステム設定項目の例

- 各種参照サーバーの設定 (NIS server, Name server)
- NFS exports, automount の設定
- ntp の設定
- メール関連の設定 (sendmail.cf)
- プリンター関連の設定 (スプールの作成、 printcap)
- tcpwraper, pam, securetty の設定

  信頼できるホスト (管理ホスト) からネットワーク経由での root による login 可に設定。
- ssh/rsh の設定

  管理サーバーからの root によるアクセスを可能に設定すること。
- system-wide な cshrc, login の設定
- システムの運用管理のための各種ソフトの設定

4.3 システムの運用管理のための各種ソフトの設定

- デスクトップの設定 (GNOME, Enlightenment の設定)

  ゲームは消去。スクリーンセーバーも消去。

- Xresources の設定; kterm, tgif, ...
- クラスター用ソフトウエア (PVM, MPI, Queue) の設定
- その他。

4.3 システムの運用管理のための各種ソフトの設定

- デスクトップの設定 (GNOME, Enlightenment の設定)
  ゲームは消去。スクリーンセーバーも消去。
- Xresources の設定; kterm, tgif, ...
- クラスター用ソフトウエア (PVM, MPI, Queue) の設定
- その他。

4.4 設置後のシステム設定

- ホストネーム、IP アドレスの設定
- デフォールトプリンターの設定

各種サーバーは役割毎に別名を設定し、 DNS において別名と本名とのマッピング (CNAME) 変更によりサーバーの変更が容易なよう計った。

各種サーバーは役割毎に別名を設定し、DNS において別名と本名とのマッピング (CNAME) 変更によりサーバーの変更が容易なよう計った。

```
ns              IN      A       133.8.40.240
...                     ...
ntp             IN      CNAME   ...
clock           IN      CNAME   ...
www             IN      CNAME   ...
proxy           IN      CNAME   ...
cache           IN      CNAME   ...
pop             IN      CNAME   ...
imap            IN      CNAME   ...
news            IN      CNAME   ...
mailhost        IN      CNAME   ...
smtp            IN      CNAME   ...
ftp             IN      CNAME   ...
dsv0            IN      CNAME   ...
dsv1            IN      CNAME   ...
dsv2            IN      CNAME   ...
dsv3            IN      CNAME   ...
```

5 運用管理

5.1 システムの運用管理

- autolog プログラムで、一定の時間入力データのないリモート接続 / デスクトップは強制終了。
- X server の終了時に、 暴走しがちな netscape 等のプログラムを kill。
- house-keeping: 管理サーバーの一つにおいて、 cron により深夜 house-keeping 用プログラムを ssh を用いリモート実行。
  - 暴走しがちなプログラムを kill。
  - xdm/gdm を再起動。
  - 各 PC 上のプリンタースプールにある未処理のファイルを消去、キューを再起動。

6　ソフトウエア構成

6.1　デスクトップとしてのソフトウエア構成

デスクトップ

　　gnome/enlightenment

ブラウザー

　　netscape

グラフィクス, DTP

　　gimp, tgif, tetex + jtex, gs, gv, acroread

ソフトウエア開発環境

　　emacs/xemacs, code-crusader, code-medic

数学

　　R, octave, scilab

数学ライブラリー

　　BLAS (Basic Linear Algebra Subprogram)

　　LAPACK (Linear Algebra Package)

　　ATLAS (Automatically Tuned Linear Algebra Software)

　　FFT (Fast Fourier Transforms)

6.2 PC クラスターとしてのソフトウエア構成

高速 IO 用の一時ファイル領域: ATA-IDE disk に 2 GB

6.2 PC クラスターとしてのソフトウエア構成

高速 IO 用の一時ファイル領域: ATA-IDE disk に 2 GB

分散メモリー並列計算用のソフトウエア

1. MPI (Massage Passing Interface); MPICH

   MPI 用の数学ライブラリーとして BLACS (Basic Linear Algebra Communications Subprogram), SCALAPACK (Scalable Linear Algebra Package)

2. PVM (Parallel Virtual Machine)

   PVM 用の数学ライブラリーとして BLACS, SCALAPACK

6.2 PC クラスターとしてのソフトウエア構成

高速 IO 用の一時ファイル領域: ATA-IDE disk に 2 GB

分散メモリー並列計算用のソフトウエア

1. MPI (Massage Passing Interface); MPICH

   MPI 用の数学ライブラリーとして BLACS (Basic Linear Algebra Communications Subprogram), SCALAPACK (Scalable Linear Algebra Package)

2. PVM (Parallel Virtual Machine)

   PVM 用の数学ライブラリーとして BLACS, SCALAPACK

Queue: Load balancing/distributed batch processing and local rsh replacement system

Queue の長所:

- ジョブ投入の優先順位、同時実行するジョブ数、 メモリー使用可能量、 CPU 時間の上限等を設定できる。
- 会話的にジョブを実行することも可能。

Queue の短所:

- load average に基づいて負荷の小さい計算機上でジョブを実行するので、ジョブ投入後 load average に反映するまで、次のジョブ投入を待つ必要がある。

6.3 クラスターとしての運用方針

- クラスターとしての使用を促進するため、各 PC へは外部から login できないよう制限し、通常のバックグランドでのジョブの実行はデスクトップ利用者以外には不可能なように管理している。
- 利用者は、研究室の計算機におけるファイルシステムを (一定の名称を用いて) automount 可能。 だだし、そのファイルシステムに関するアクセス制限は、利用者の責任である。

6.4 Queue の使用例

Queue を用いて、サーバーよりジョブを投入した例を下に示す。

```
ug% queue -q -d days -n -w -- sol_4_L60.sh 3.570 </dev/null >& sol_4_L60.log &
```

6.4 Queue の使用例

Queue を用いて、サーバーよりジョブを投入した例を下に示す。

```
ug% queue -q -d days -n -w -- sol_4_L60.sh 3.570 </dev/null >& sol_4_L60.log &

ug% ps fxa
26012 ?        S      0:00 queue -q -d days -n -w -h ug -- SMB25L120M05M1MN100.sh
26169 ?        S      0:00 queue -q -d days -n -w -- SMB12L100M05M1MN80.sh
26191 ?        S      0:00 queue -q -d days -n -w -- SMB10L60M05M1MN50.sh
26203 ?        S      0:00 queue -q -d days -n -w -- SMB9L60M05M1MN40.sh
26206 ?        S      0:00 queue -q -d days -n -w -- SMB8L60M05M1MN30.sh
26208 ?        S      0:00 queue -q -d days -n -w -- SMB7L60M05M1MN20.sh
...
...
...
  913 ?        S      0:00 queue -q -d days -n -w -- sol_4_L60.sh 3.514
  968 ?        S      0:00 queue -q -d days -n -w -- sol_4_L60.sh 3.568
 1006 ?        S      0:00 queue -q -d days -n -w -- sol_4_L60.sh 3.570
ug%
```

6.4 Queue の使用例

Queue を用いて、サーバーよりジョブを投入した例を下に示す。

```
ug% queue -q -d days -n -w -- sol_4_L60.sh 3.570 </dev/null >& sol_4_L60.log &

ug% ps fxa
26012 ?        S      0:00 queue -q -d days -n -w -h ug -- SMB25L120M05M1MN100.sh
26169 ?        S      0:00 queue -q -d days -n -w -- SMB12L100M05M1MN80.sh
26191 ?        S      0:00 queue -q -d days -n -w -- SMB10L60M05M1MN50.sh
26203 ?        S      0:00 queue -q -d days -n -w -- SMB9L60M05M1MN40.sh
26206 ?        S      0:00 queue -q -d days -n -w -- SMB8L60M05M1MN30.sh
26208 ?        S      0:00 queue -q -d days -n -w -- SMB7L60M05M1MN20.sh
...
...
...
  913 ?        S      0:00 queue -q -d days -n -w -- sol_4_L60.sh 3.514
  968 ?        S      0:00 queue -q -d days -n -w -- sol_4_L60.sh 3.568
 1006 ?        S      0:00 queue -q -d days -n -w -- sol_4_L60.sh 3.570
ug%

ug% kill -9 1006
```

7　ネットワークセキュリテー

- rsh/rlogin を使用せず ssh/telnet-ssl を用いている。
- 学生のパスワード等は秘密保持不可能であるという前提で、 PC から学内へのアクセスは、学外の計算機と同じ扱いをしている。
- PC から学外へはアクセスできるが、学外からはアクセスできないようにルーターで設定している。
- 学外からサーバーへのログインは、例外を除き、 tcpwrapper および pam の設定で、
  - 学部学生の場合は、禁止。
  - 大学院学生の場合は、国内の大学、政府研究機関のみ可。
  - 教官の場合は国内外の大学、研究機関のみ可能。

8 ま と め

教育用として使用する 120 台の PC をクラスター構成により研究システムとして共用するシステムを紹介した。 Celeron 500MHz は最新の Pentium III の半分以下の CPU 能力と思われるが、クラスター全体のジョブ処理能力は 120/3 = 40 倍である。パラメーターを変えて多数回実行する必要のあるジョブの場合大変有用である。

参考資料

1. http://www.debian.org/
2. http://www.gnu.org/


群馬大学工学部
記号
番号
B56R11−1−45−1
12・3・24
NisseiCom

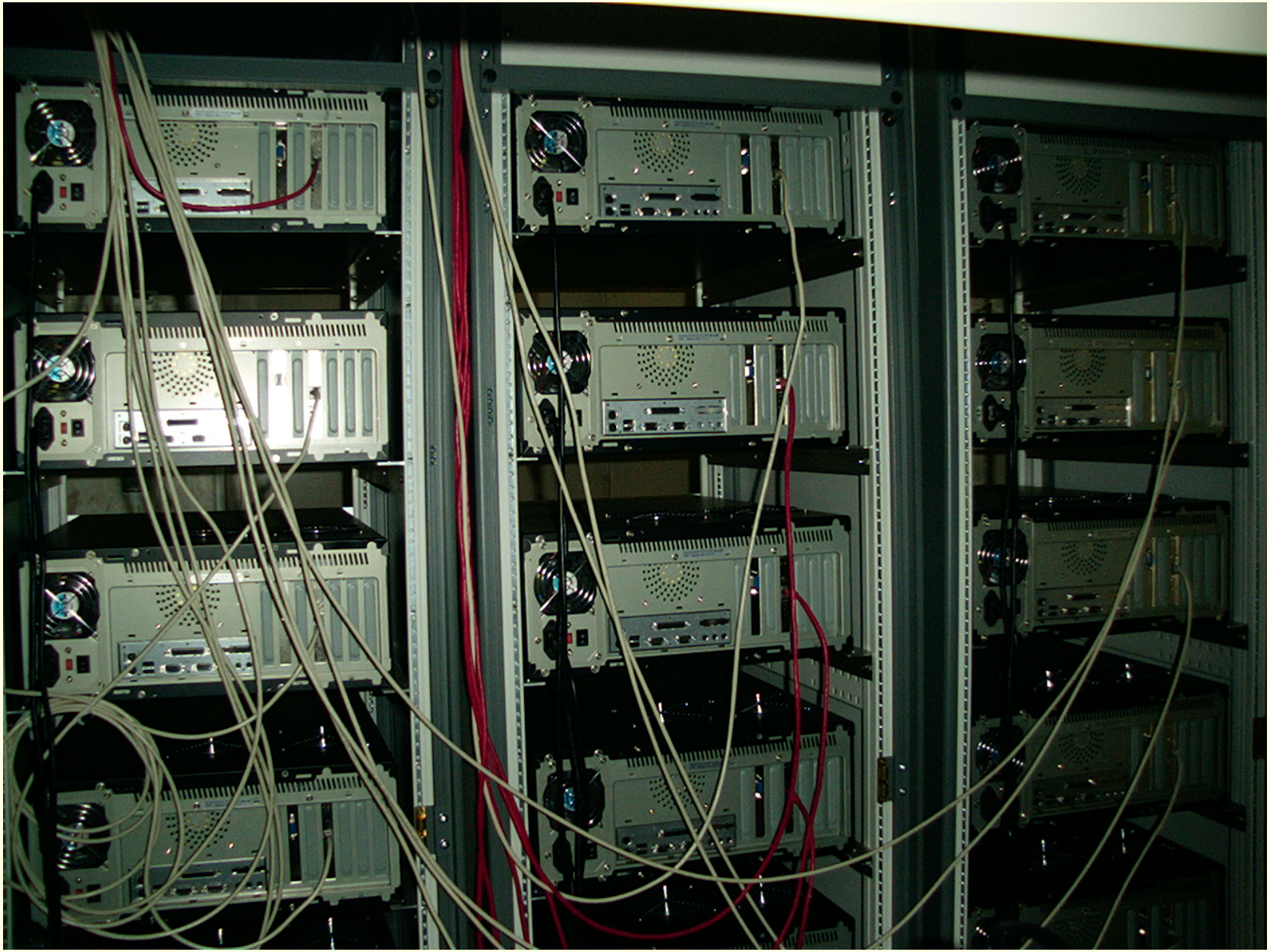


表 1: ソフトウエア構成

Linux	debian potato
Desktop	gnome/enlightenment
Graphics	gimp
Editors	emacs/xemacs, code-crusader, code-medic
DTP	tetex + jtex, gs, gv, acroread
Script languages	awk, perl, python, tcl/tk, scheme
Mathematics	R, octave, scilab
Mathematical libraries	BLAS/ATLAS, LAPACK, FFT
Libraries for parallel comp.	PVM, (BLACS, SCALAPACK)
	MPICH MPI (BLACS, SCALAPACK)
	LAM MPI (BLACS, SCALAPACK)
Queueing for cluster	queue, DQS
Temporary file space for better I/O	2 GB in local ATA-IDE disk
Secure ...	telnet-ssl, open ssh

BLAS: Basic Linear Algebra Subprogram

ATLAS: Automatically Tuned Linear Algebra Software

LAPACK: Linear Algebra Package

FFT: Fast Fourier Transforms

BLACS: Basic Linear Algebra Communications Subprogram

SCALAPACK: Scalable Linear Algebra Package

PVM: Parallel Virtual Machine

MPI: Massage Passing Interface

Queue: Load balancing/distributed batch processing and local rsh replacement system

DQS: Distributed Queueing System