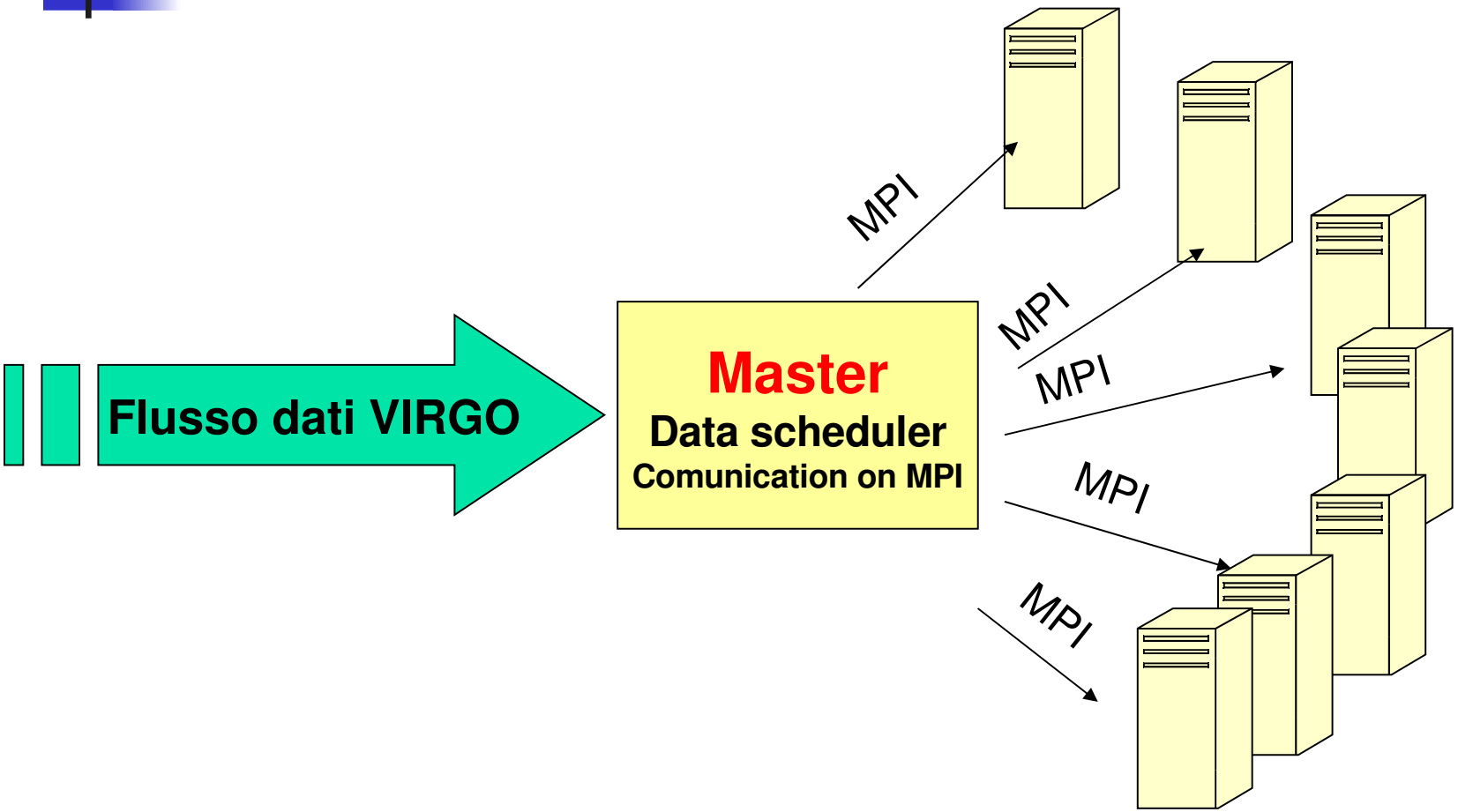
VIRGO

- **Virgo:**
  - **Interferometro in fase di ultimazione:**
    - **prevista accensione inizi 2003**
  - **La parte centrale dell'interferometro già attiva ed in presa dati**
    - **Già effettuati 3 engineering runs, il quarto la settimana prossima**
    - **1,5TB di dati per ciascun run**

Schema di rivelazione

- Il modello di rivelazione di Virgo mal si combina con non un sistema classico di *batch machine*
- Il nostro schema di *detection* necessita un sistema distribuito (quasi parallelo) di calcolo tramite opportuna libreria di message passing (MPI)
- Lo stesso stream di dati subisce in parallelo il medesimo algoritmo di *detection* con parametri diversi

Schema di analisi dati..



VIRGO-PG Beowulf

Un grosso (300GFlops) cluster è previsto in Virgo

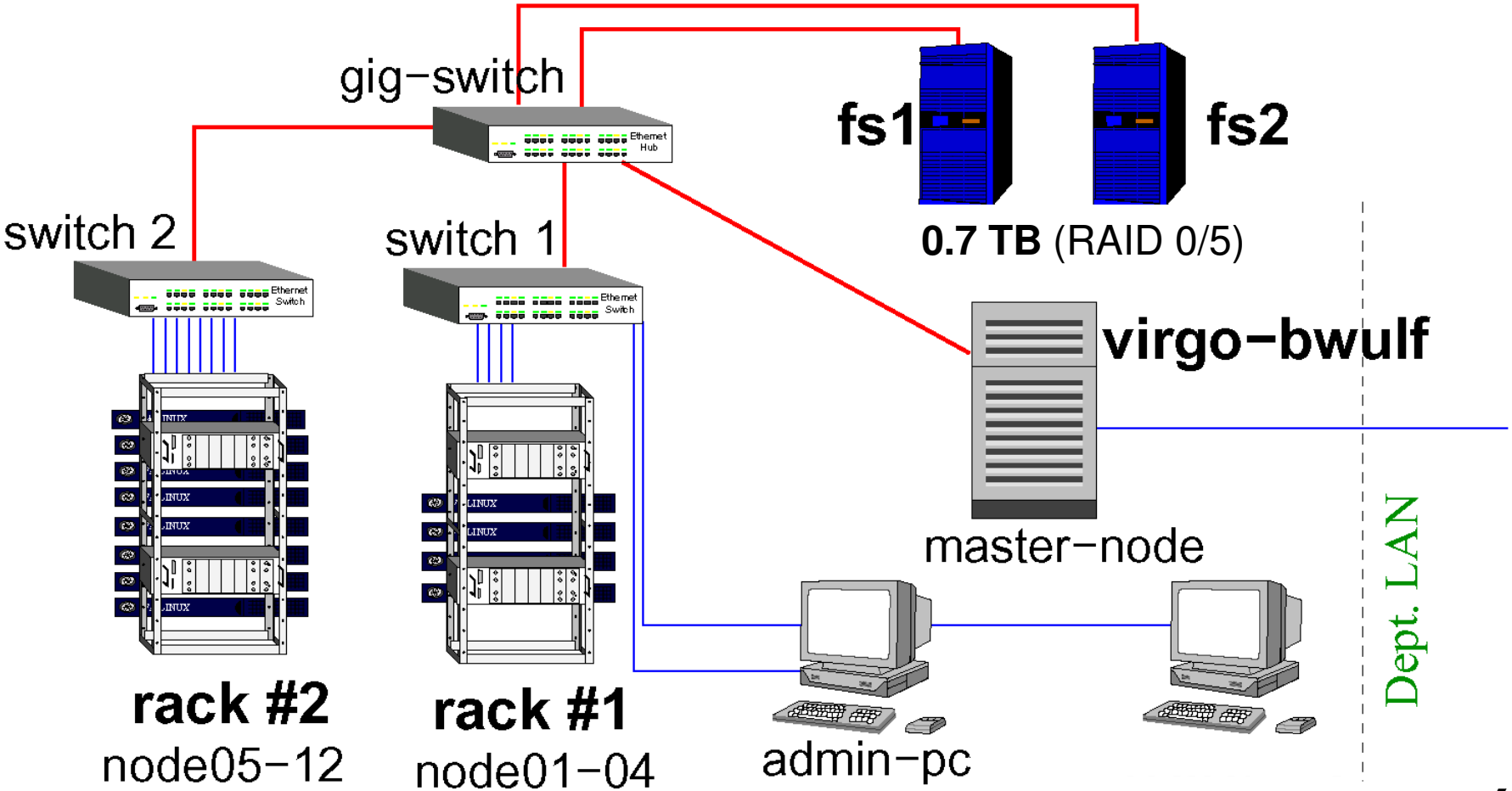
Alcune farms di studio sono state sviluppate nelle diverse sezioni

Cluster di Perugia

Obiettivi:

- Configurazione
- Performances
- Scalabilità
- Gestione
- Software di detection e ricostruzione
- Test per l'on-line di Virgo

VIRGO-PG Beowulf (II)



Specifiche Hardware

Nodi di calcolo(24 CPUs):

- 4 nodes with dual Pentium-III 866 MHz CPUs, 512 Mb RAM, 2 Fast-Ethernet NICs
- 8 nodes with dual Pentium-III 1 GHz CPUs, 1 Gb RAM, 2 Fast-Ethernet NICs

Master node:

- dual Athlon 1 GHz CPUs, 512 Mb RAM, 2 Fast-Ethernet NICs, Gigabit-Ethernet NIC

Fileserver 1 (360 Gb):

- dual Pentium-III 1 GHz, 512 Mb RAM, Fast-Ethernet NIC, Gigabit-Ethernet NIC
- Promise RAID-0 IDE controller
- IDE ultra-DMA disks

Fileserver 2 (360 Gb):

- dual Pentium-III 1 GHz, 256 Mb RAM, Fast-Ethernet NIC, Gigabit-Ethernet NIC
- Adaptec RAID-5 SCSI controller
- SCSI ultra-160 disks

Software setup (I)

``Beowulf'' or ``MOSIX'' operation

Approccio Single System Image:

- punto di controllo singolo dell'immagine kernel e sequenza di boot per tutti i nodi
- punto di controllo singolo delle librerie di sistema
- punto di controllo singolo delle librerie delle applicazioni (/beowulf)
- punto di controllo singolo dell'user environment (/beowulf/env)
- i nodi condividono un'identico namespace

Principi Guida:

- Seguire una filosofia Open!
  - *GNU/Linux operating system, distribuzione Debian*
  - *Tool e librerie Free Software e Open Source*
- Mantenere il sistema semplice!
  - *Nessun daemon o servizio non necessario*
  - *No NIS/NIS+*
  - *Nessuna immagine di sistema replicata*
  - *Debian advanced packaging system*

Software setup (II)

Master node:

- login server (SSH), gateway verso la rete esterna, NAT, firewall
- macchina di sviluppo
- funzioni di controllo e monitoring
- NFS server (/beowulf filesystem, NFS root filesystem per operazioni diskless)
- DHCP e TFTP server
- NTP (time) server
- log server

Fileserver 1, fileserver 2:

- NFS server

Nodi di calcolo:

- COMPUTATION
- disk-based o operazioni disk-less
- Parallel Virtual Filesystem (PVFS)
- facilmente e dinamicamente riconfigurabili

Librerie e tools

Compilatori:

- GNU suite: GCC, G++, G77, GCJ
- Intel C/C++ compiler, Intel Fortran77/90 compiler

Librerie di Message-Passing:

- MPI/LAM
- MPI/MPICH
- PVM

Librerie Scientifiche:

- FFTW
- ATLAS
- BLAS
- LAPACK
- SCALAPACK
- Siglib
- VIRGO libraries (Framelib)

Principi di installazione

Motivazione di costi

OSCAR è troppo semplice!

Approccio alla clusterizzazione Single System Image (SSI)

- **fornire agli utenti un'immagine del cluster unificata**
- **richiede un'integrazione a livelli multipli: kernel, filesystem, namespace, . . .**
- **. . . ma non esiste ancora!**

Il root filesystem singolo è il primo passo verso l'SSI:

Maneggiabilità

Flessibilità

A volte, quando poi avete dischi, non volete toccarli

Installazione nodi diskless

- **Dove eseguire il boot?**
  - ***Floppy***
  - ***Network (PXE, etherboot)***
- **Dove risiede il mio root filesystem?**
  - ***Network file system (NFS)***
- **Dove eseguire lo swap?**
  - ***Nessuno swap.***
  - ***Swap su rete***
  - ***Swap su disco***

In più, se si desidera una filesystem root singolo per tutti i nodi:

- **Dove risiede l'identità di ciascun nodo?**
  - ***Creata on the fly***
  - ***Memorizzata remotamente***

Le nostre scelte

- **Root filesystem su NFS, montato read-only**
  - *root filesystem basato su distribuzione Debian*
  - *root filesystem esportato verso i nodi è mantenuto utilizzando i tool standard Debian(dpkg; apt-get)*
- **GRUB bootloader**
  - *Integra un client BOOTP/DHCP*
  - *Può scaricare ed eseguire script, specificando una sequenza di boot*
  - *Il boot è eseguito tramite floppy*
- **Piccola ram-disk, per accessi in scrittura volatili**
  - /tmp deve essere scrivibile*! (pensate ai file di lock . . .)*
- **Devfs virtual filesystem**
  - *Nessun device inode lookups over the network*
  - *Nessun file* /dev *è presente nel filesystem root esportato*
- **NFS-shared /beowulf filesystem**
  - *Contiene le librerie delle applicazioni (MPICH, LAM, FFTW, . . .)*
  - *Contiene le home directory*

Un'occhiata ad un nodo..

```
cattuto@node10:~$ df -aTm
Filesystem        Type   1M-blocks   Used Available Use% Mounted on
/dev/root          nfs        5613   1341      3987  26% /
none               devfs         0      0         0   -  /dev
proc               proc          0      0         0   -  /proc
/dev/ram0          ext2          8      1         7   1% /ramdisk
node00:/beowulf    nfs       10199   8524      1676  84% /beowulf
```

```
cattuto@node10:~$ ls -l /tmp
lrwxrwxrwx  1 root  root    12 Jan 30 13:48 /tmp -> /ramdisk/tmp/
```

```
cattuto@node10:~$ ls -l /ramdisk
drwxr-xr-x  3 root  root  1024 Feb  8 15:02 etc/
drwxrwxrwt  5 root  root  1024 Feb 14 00:17 tmp/
drwxr-xr-x  9 root  root  1024 Feb  6 18:08 var/
cattuto@node10:~$
```

Configurazione del kernel (I)

```
Linux Kernel v2.4.17 Configuration

                    Networking options
  Arrow keys navigate the menu.  <Enter> selects submenus --->.
  Highlighted letters are hotkeys.  Pressing <Y> includes, <N> excludes,
  <M> modularizes features.  Press <Esc><Esc> to exit, <?> for Help.
  Legend: [*] built-in  [ ] excluded  <M> module  < > module capable
        ^(-)
        <*> Unix domain sockets
        [*] TCP/IP networking
        [ ]   IP: multicasting
        [ ]   IP: advanced router
        [*]   IP: kernel level autoconfiguration
        [*]     IP: DHCP support (NEW)
        [*]     IP: BOOTP support (NEW)
        [ ]     IP: RARP support (NEW)
        < >   IP: tunneling
        < >   IP: GRE tunnels over IP
        v(+)

                <Select>    < Exit >    < Help >
```

Abilitare la configurazione automatica IP del kernel in fase di boot

Configurazione del kernel (II)

```
Linux Kernel v2.4.17 Configuration

                    Network File Systems
  Arrow keys navigate the menu.  <Enter> selects submenus --->.
  Highlighted letters are hotkeys.  Pressing <Y> includes, <N> excludes,
  <M> modularizes features.  Press <Esc><Esc> to exit, <?> for Help.
  Legend: [*] built-in  [ ] excluded  <M> module  < > module capable

      < > Coda file system support (advanced network fs)
      <*> NFS file system support
      [*]   Provide NFSv3 client support
      [*]   Root file system on NFS
      < > NFS server support
      < > SMB file system support (to mount Windows shares etc.)
      < > NCP file system support (to mount NetWare volumes)

                <Select>    < Exit >    < Help >
```

Abilitare NFS ed il supporto NFSboot

Configurazione del kernel (III)

```
Linux Kernel v2.4.17 Configuration

                          Block devices
  Arrow keys navigate the menu.  <Enter> selects submenus --->.
  Highlighted letters are hotkeys.  Pressing <Y> includes, <N> excludes,
  <M> modularizes features.  Press <Esc><Esc> to exit, <?> for Help.
  Legend: [*] built-in  [ ] excluded  <M> module  < > module capable

        <*> Normal PC floppy disk support
        < > XT hard disk support
        < > Compaq SMART2 support
        < > Compaq Smart Array 5xxx support
        < > Mylex DAC960/DAC1100 PCI RAID Controller support
        <M> Loopback device support
        <M> Network block device support
        <*> RAM disk support
        (4096)   Default RAM disk size
        [ ]   Initial RAM disk (initrd) support

                <Select>    < Exit >    < Help >
```

Abilitare il supporto RamDisk

La sequenza di boot

1. Power on. GRUB viene caricato da dischetto (o ROM PXE)
2. GRUB identifica la periferica di rete NIC, ed invia una richiesta DHCP per configurare la rete
3. GRUB download (via TFTP) uno script di boot dal server DHCP
4. Lo script di boot preleva un'immagine kernel e l'avvia con gli opportuni parametri
5. Il kernel linux esegue il boot ed il codice di auto configurazione IP invia una richiesta DHCP per configurare la rete
6. Viene montato il filesystem root su NFS(read-only) dal dhcp server
7. Inizia la configurazione di sistema.
8. Devfs è montato
9. La ramdisk è creata, popolata e montata
10. Lo script di boot completa la configurazione locale(su ramdisk)
11. Il filesystem remoto /beowulf viene montato
12. Viene effettuata una configurazione specifica per nodo(se esiste)
13. Il sistema è pronto!

Uno sguardo al lato server..

```
virgo-bwulf:/beowulf/boot# ls -l
  1 root  root 827361 Nov 23 16:18 bzImage-2.4.17
  1 root  root 921385 Nov 18 16:13 bzImage-2.4.17-mosix
  1 root  root    151 Jan 29 21:34 install.lst
  1 root  root    232 Jan 29 21:34 local.lst
  1 root  root    139 Jan 29 21:34 node-rw.lst
  1 root  root    102 Jan 29 21:34 node.lst
  1 root  root     11 Feb 14 00:49 node10.lst -> install.lst
```

- **Al momento del boot, GRUB configura la rete tramite DHCP**
- **Successivamente, ogni nodo esegue il download ed esegue uno script di boot GRUB:**
  - ***Se esiste uno script specifico per il nodo, esegue quello***
  - ***altrimenti, utilizza il generico* node.lst**
- **Lo script di boot preleva il kernel giusto e lo avvia**
  - ***La command line del kernel può essere manipolata da GRUB***
- **Il kernel e la sequenza di boot può essere controllata per ciascun nodo, cambiando**

  i file ed i link simbolici in /beowulf/boot – e solo questi.

Una giornata per montaggio e cablaggio…

5 minuti per inizializzare il sistema e renderlo operativo...

- **Uno sguardo al file di log.....**

Prima richiesta DHCP

```
==========================================================================
(POWER-ON DEI NUOVI NODI, PRIVI DI SISTEMA OPERATIVO)
Jan 30 16:42:20 virgo-bwulf dhcpd-2.2.x: DHCPDISCOVER from 00:20:ed:11:27:e8 via eth1
Jan 30 16:42:20 virgo-bwulf dhcpd-2.2.x: DHCPOFFER on 192.168.1.105 to 00:20:ed:11:27:e8 via eth1
Jan 30 16:42:22 virgo-bwulf dhcpd-2.2.x: DHCPDISCOVER from 00:20:ed:11:09:1a via eth1
Jan 30 16:42:22 virgo-bwulf dhcpd-2.2.x: DHCPOFFER on 192.168.1.107 to 00:20:ed:11:09:1a via eth1

.................

Jan 30 16:45:31 node11 kernel: IP-Config: Complete:
Jan 30 16:45:31 node11 kernel:      device=eth1, addr=192.168.1.111, mask=255.255.255.0, gw=192.168.1.1,
Jan 30 16:45:31 node11 kernel:     host=node11, domain=cluster, nis-domain=(none),
Jan 30 16:45:31 node11 kernel: Freeing unused kernel memory: 224k freed
(SISTEMA INSTALLATO E CONFIGURATO SU 8 NODI, 16 NUOVE CPU ONLINE)
==========================================================================
```

Ultimo messaggio di boot

Pro e contro

Pro:

- **Maneggiabilità**
  - punto di controllo singolo dell'immagine kernel e sequenza di boot per tuttii nodi
  - punto di controllo singolo delle librerie di sistema
  - punto di controllo singolo delle librerie delle applicazioni (/beowulf)
  - punto di controllo singolo dell'user environment (/beowulf/env)
  - i nodi di calcolo condividono un'identico namespace
  - operazioni chrooted sul filesystem dei nodi, utilizzando
- **Flessibilità**
  - I nodi non necessitano di avere lo stesso set di daemon
  - I dischi locali, se presenti, possono essere partizionati e popolati automaticamente (utili per swap locale, PVFS, …)
- **Nessuna perdita di performance dovuta ad istallazione locale**

Contro:

- **Scalabilità**
  - Durante la fase di boot, tutti i nodi accedono alla stesso filesystem NFS (dove il sistema di caching VFS alleggerisce gli accessi)
  - In pratica: non è stato osservato nessun rallentamento nella sequenza di boot

Referenze

- GRUB

http://www.gnu.org/software/grub/

- Linux NFS-Root mini-HOWTO

http://www.linuxdoc.org/HOWTO/mini/NFS-Root.html

- Linux Diskless HOWTO

http://www.linuxdoc.org/HOWTO/Diskless-HOWTO.html

- Devfs FAQs

http://www.atnf.csiro.au/rgooch/linux/docs/devfs.html

- Debian

http://www.debian.org/

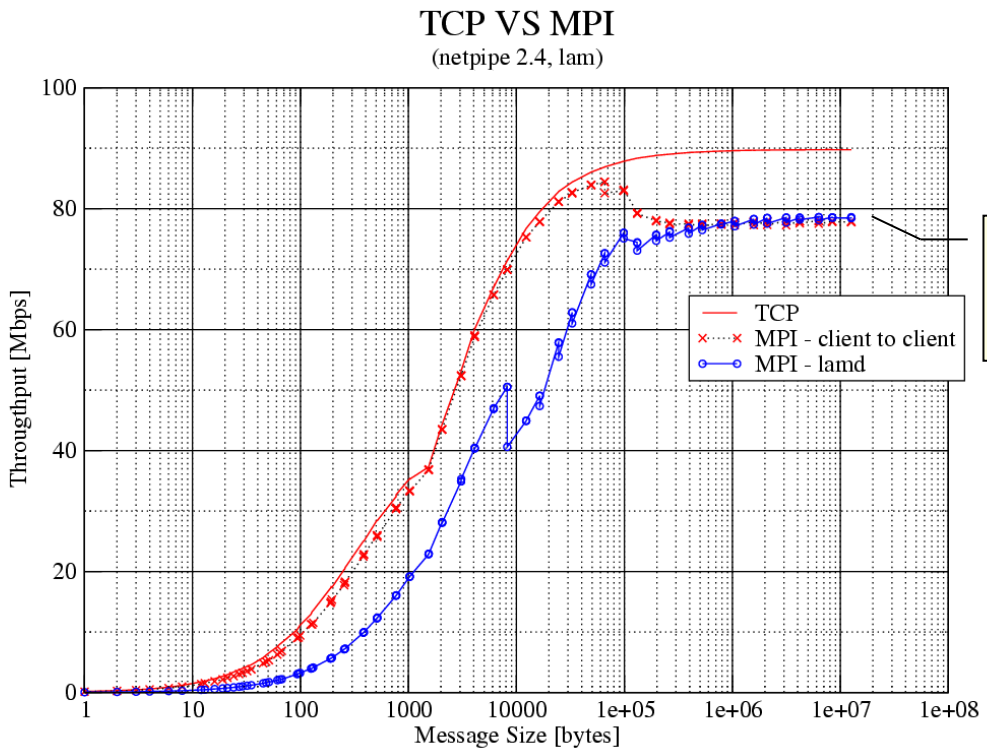
- Single System Image Clusters for Linux

http://ssic-linux.sourceforge.net/

- Scyld

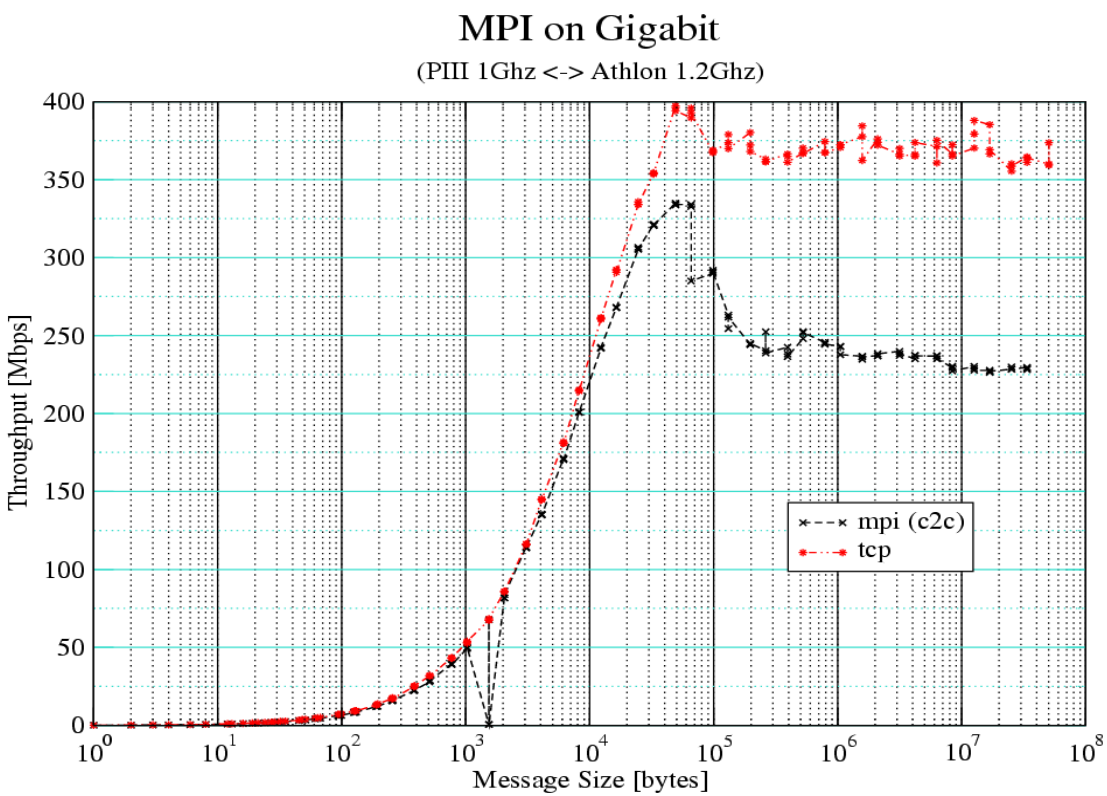
http://www.scyld.com/

FastEthernet – MPI: Test di comunicazione



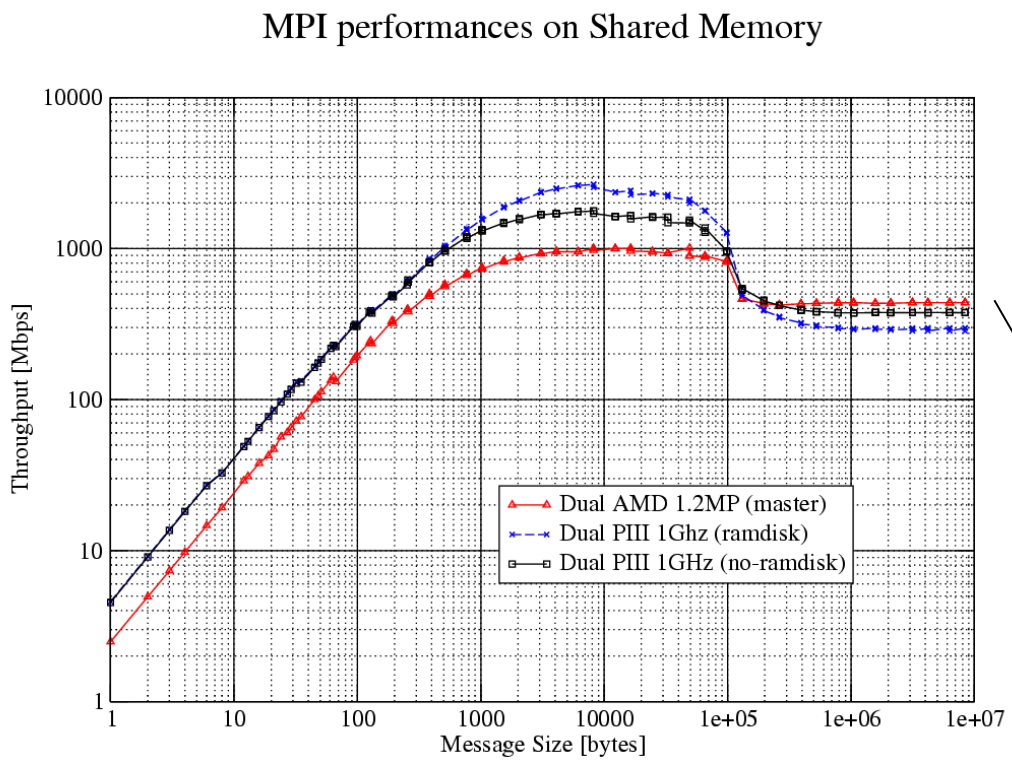
Overhead di comunicazione MPI

Gigabit – MPI: throughput in RETE



- Throughput:
  - Massimo: 400Mbps
  - A regime: 370Mbps
- Aggiornamento driver Intel da ver 3.x a 4.x:
  - +5% throughput
  - Buffering ridotto
- Collo di bottiglia
  - Switch- MTU
- Soluzione
  - **Jumbo Frame**

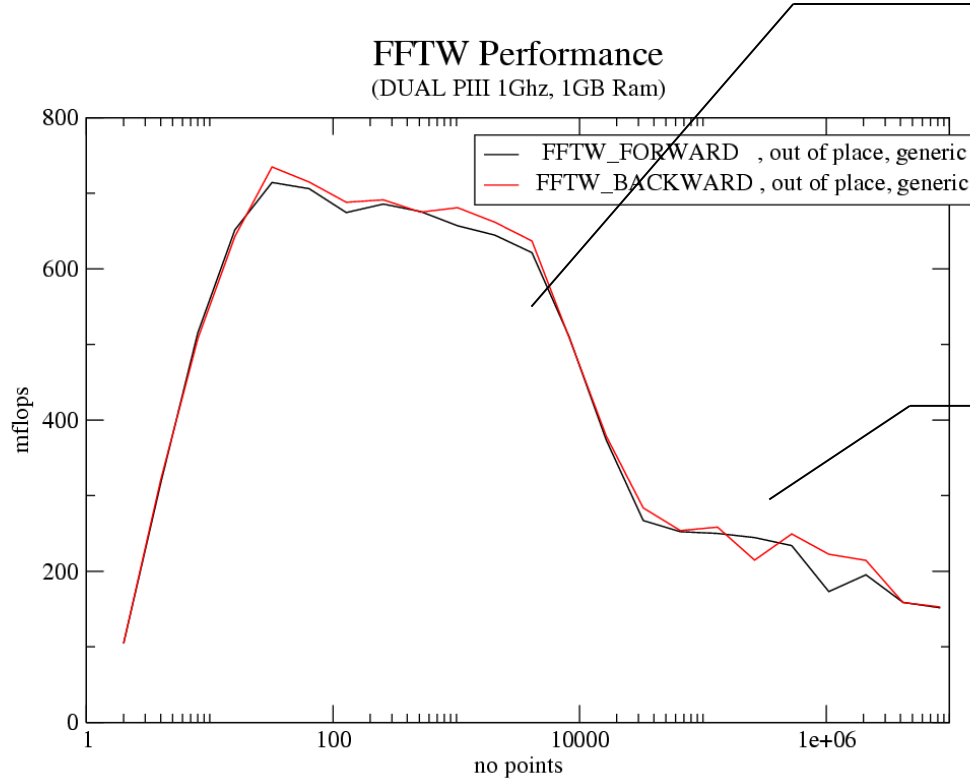
MPI in shared memory: throughput in RAM



- Throughput:
  - Cache
    - Massimo: **3G**bps
    - Regime: **2G**bps
  - Out-cache
    - **300-400 M**bps

Limitati in banda passante RAM

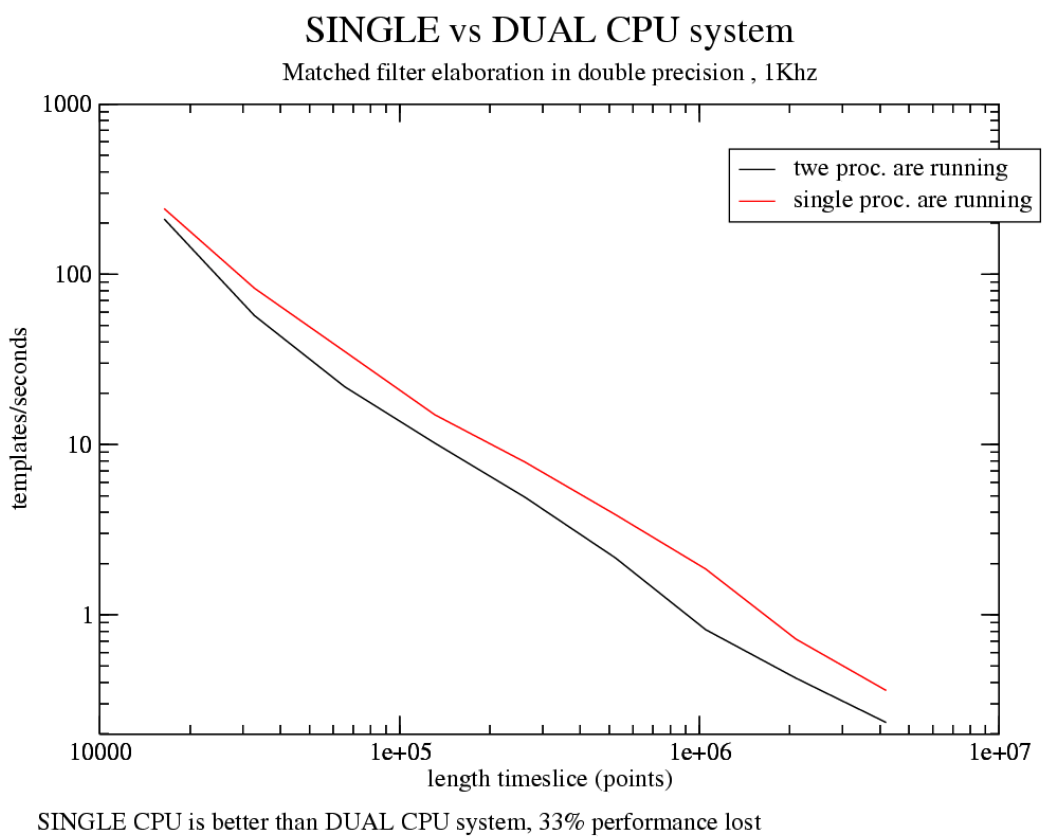
FFTw: Operare in cache / ram



Operazioni in Cache

Operazioni in RAM

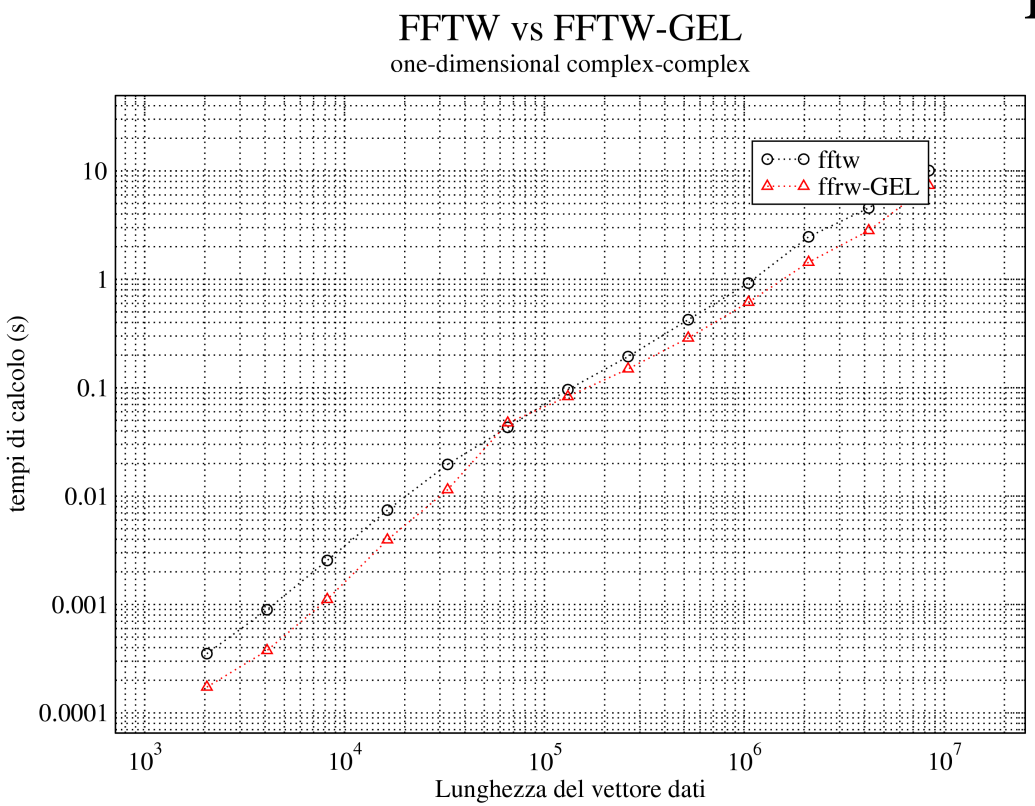
..sistemi dual... Perdita di prestazioni



- Calcoli con elevato IO-Ram:

Perdita del
30%
nei Sistemi Dual PIII

..prima soluzione..
CPU dedicata: FFTW-GEL



Librerie dedicate al tipo di CPU utilizzata:

- **AMD: Athlon**
- **Intel: PIII, PIV**

Prestazioni:

+10-40%

Ref: http://www.fftw.org

...soluzione AMD... DUAL Athlon

Sistemi DUAL testati	MB/s
PIII 1GHz (lowcost MotherBoard)	331
PIII 1GHz (no-so-lowcost MB - ECC Ram)	401
PIII 866MHz (no-lowcost MB - ECC Ram)	419
Athlon 1.2MP (..ma è un forno..)	**815.1**

Test eseguito con stream:
http://www.cs.virginia.edu/stream/

Ringraziamenti

- VIRGO group
- Leone Bosi (leone.bosi@pg.infn.it)
- Ciro Cattuto (ciro.cattuto@pg.infn.it)