IBM Scalable POWERparallel System 2

Vulcan

Giuseppe Vitillaro

Dipartimento di Chimica

Universita’ degli Studi di Perugia

e-mail: <peppe@unipg.it>

Perugia, 10 Ottobre 1996

La macchina

- l'SP2 e' una macchina parallela a memoria distribuita
- consiste di un numero variabile di nodi (da 2 a 128) organizzati in frames
- ogni frame puo' contenere da 2 a 16 nodi
- i singoli nodi sono processori di architettura IBM RISC/6000 POWER2
- frames e nodi possono essere interconnessi fra loro mediante uno «switch» ad alta velocita' e bassa latenza, l'High Perfomance Switch (HPS)

La macchina

- Esistono diversi tipi di nodi. Alcuni esempi:
  - **Thin** (equivalente ad un RISC/6000 390)

    IBM RISC/6000 POWER2 clock 66.7Mhz

    32Kb I-cache 64K D-cache

    4 slots Micro Channel 80Mb/sec 64 bits data bus
  - **Thin 2** (equivalente ad un RISC/6000 390H)

    IBM RISC/6000 POWER2 clock 66.7 Mhz

    32Kb I-cache 64Kb D-cache

    4 slots Micro Channel 80Mb/sec 64 bits data bus
  - **Wide** (equivalente ad un RISC/6000 590)

    IBM RISC/6000 POWER2 clock 66.7Mhz

    32Kb I-cache 256Kb D-cache

    8 slots Micro Channel 80Mb/sec 256 bits data bus

La macchina

- I frames sono suddivisi in «drawers», ognuno dei quali puo' contenere 2 nodi thin o un nodo wide

- ciascun frame puo' ospitare 8 drawers, per un totale di 16 nodi thin, 8 nodi wide o una combinazione di nodi thin e wide

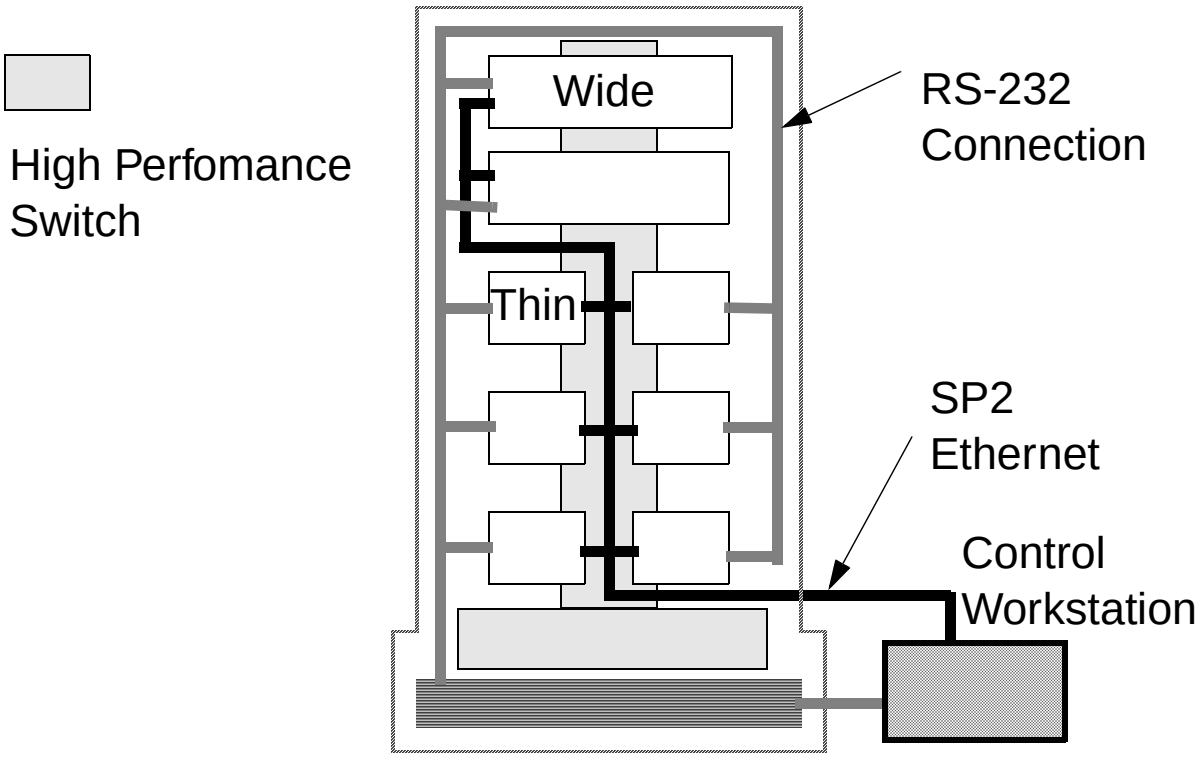
- i frames possono essere interconnessi mediante l'HPS e raggiungere (nei modelli piu' espandibili dell'SP2) centinaia di nodi.

- L'SP2 installato a Perugia e' un

IBM 9076 SP2 modello 302

e puo' collegare 80 nodi

La macchina


High Perfomance
Switch
Wide
Thin
RS-232
Connection
SP2
Ethernet
Control
Workstation

I nodi

- I nodi condividono architettura hardware e software con la piattaforma IBM RISC/6000 POWER2
- Sulla piastra madre di ciascun nodo sono integrati una scheda di rete Ethernet e una scheda SCSI-2 Fast/Wide
- Su ogni nodo (se e' presente lo switch) puo' essere installato un HPS Adapter-2 (i modelli HPS Adapter-1 sono ormai obsoleti) che lo connette all'HPS
- Sui nodi thin possono essere installati da 1 a 9 gigabytes di disco locale e da 64 a 512Mb di memoria centrale (i nodi wide sono piu' espandibili)

I nodi

- Il sistema operativo dei nodi e' AIX/6000 3.2.5
- su ciascun nodo gira una copia indipendente di AIX/6000
- Il sistema operativo puo' essere caricato da un disco locale o via rete
- i nodi possono comunicare sia via Ethernet che via HPS
- una parte della macchina e' dedicata ad acquisire informazioni dai nodi attraverso linee seriali asincrone RS-232

La Control Workstation

- La Control Workstation e’ una workstation RISC/6000 che svolge funzioni di controllo e di servizio
- E’ connessa attraverso una linea seriale ad ogni frame e fa parte della rete Ethernet dell’SP2
- Permette di:
  - installare e configurare l’SP2
  - eseguire bootstrap e shutdown della macchina
  - monitorare l’hardware
  - svolgere funzioni di file server

High Perfomance Switch

- L'High Perfomance Switch e' il componente hardware dell'SP2 che interconnette i nodi in una rete di comunicazione ad alta velocita' e bassa latenza

- l'HPS e' un «multi-stage packed switched Omega switch» che fornisce una banda di 40Mb/sec e una latenza di 40 microsecondi, fra ogni coppia di nodi

- Su ogni nodo e' installata una scheda HPS Adapter-2, che collega il nodo all'HPS

- Gli HPS dei vari frame possono essere collegati per formare un'unica rete

High Perfomance Switch

- Le applicazioni possono comunicare attraverso lo switch in due modi radicalmente diversi:
  - ad ogni HPS Adapter-2 e' assegnato un indirizzo IP e le applicazioni possono comunicare via «socket», utilizzando i meccanismi standard TCP/IP
  - utilizzando lo «user space communication mode»: le applicazioni utilizzano una speciale libreria applicativa per comunicare direttamente attraverso lo switch

HPS: IP mode

- Via TCP/IP, piu' applicazioni possono utilizzare concorrentemente lo switch senza alcuna restrizione
- Tutto il software basato sui «Berkeley Sockets» puo' funzionare senza alcun cambiamento (NFS, AFS, PVM, Linda, Express, etc,)
- l'IP mode puo' essere penalizzante in termini di perfomance: per essere scambiati i dati devono «attraversare» un gran numero di «strati software» del sistema operativo
- Piu' utenti e piu' applicazioni possono utilizzare contemporanemante il modo IP

HPS: User Space mode

- Lo «user space communication» mode e' stato pensato per essere utilizzato da applicazioni che che richiedono elevate perfomance di comunicazione: banda elevata e bassa latenza

- Ogni nodo consente ad una sola applicazione alla volta di utilizzare lo «user space mode, concorrentemente ad applicazioni IP mode.

- E' «necessario» utilizzare una libreria applicativa proprietaria: l'applicazione deve essere stata scritta in modo da chiamare le funzioni primitive della libreria

- In «user space mode» si possono raggiungere le perfomance piu' elevate: fino a 40Mb/sec di banda e 40 microsecondi di latenza

HPS: IP verso User Space

User space mode (MPL)

Nodo	Latenza	Pt to Pt bw
Thin	40.0 mics	35.4 Mb/s
Thin 2	39.0 mics	35.7 Mb/s
Wide	39.2 mics	35.6 Mb/s

udp/IP mode (MPL)

Nodo	Latenza	Pt to Pt bw
Thin	312.1 mics	9.9Mb/s
Thin 2	270.4 mics	12.0Mb/s
Wide	268.8 mics	12.1Mb/s

Vulcan

- Presso il Centro di Calcolo dell'Universita' degli Studi di Perugia e' installato un IBM 9076 SP2 **modello 302**.

- La macchina (il cui nome e' «**vulcan**») e' costituita da un frame contenente **«8 nodi Thin»** e un High Perfomance Switch.

- I singoli nodi sono numerati da 01 a 08 ed hanno come hostname sp01,sp02,...,sp08

- Sui nodi 01,...,06 sono installati 128Mb di RAM (due schede da 64Mb) e un disco locale SCSI-2 (2Gb)

Vulcan

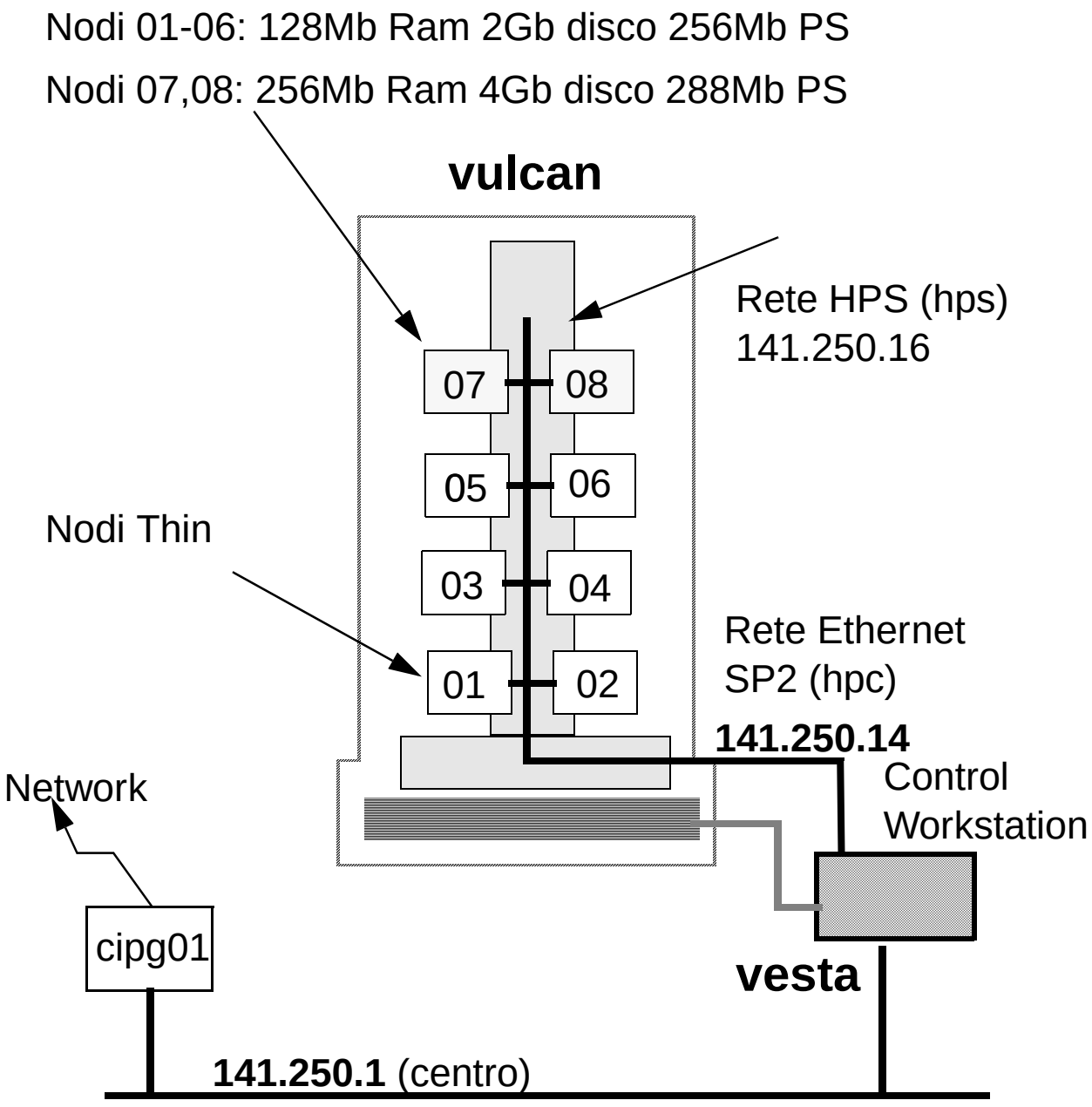
- Sui nodi 07,08 sono installati 256Mb di RAM (2 schede da 128Mb) e 2 dischi locali SCSI-2 locali (2x2Gb)

- La Control Workstation e' un RISC/6000 7011-25T (il cui hostname e' vesta) su cui sono installati 64Mb di RAM e due dischi SCSI-2 (2x2Gb)

- Su tutti i nodi e' installata una copia completa di AIX/6000 3.2.5.1 ed e' definito uno spazio di paginazione di almeno 256Mb

- La rete Ethernet dell'SP2 ha indirizzo 141.250.14. L'SP2, i suoi nodi e la Control Workstation appartengono al dominio DNS **«hpc.unipg.it»**

Vulcan



Rete del Centro di Ateneo per i Servizi Informatici

Vulcan

- La Control Workstation «vesta.unipg.it» o «vulcan.unipg.it» svolge funzioni di «router» verso la rete Ethernet dell'SP2

- Su tutti i nodi sp01,...,sp08 sono definite le interfacce di rete corrispondenti agli adapters HPS

- Le interfacce di rete HPS sono accessibili con i nomi «**sw01,sw02,...,sw08**», mentre le interfacce Ethernet coincidono con gli hostnames «**sp01,sp02,...,sp08**» (rete hpc)

- La rete HPS e' accessibile solo internamente all'SP2: non esiste routing verso l'esterno per la rete (hps) 141.250.16 (neppure verso la CW)

Vulcan

- Le home directories degli utenti sono allocate sul nodo **sp01** e sono accedute dagli altri nodi mediante **NFS** attraverso la rete HPS. La CW accede alle home directories via Ethernet.

- Su tutti i nodi e' definito un file system locale "**/u0**" di circa 800Mb e sui nodi sp07,sp08 un ulteriore file system locale "**/u1**" di circa 1.7Gb

- Da qualunque nodo e' possibile accedere uno dei file systems usando il path "**/sp/sp0x/uy**" oppure "**/sp/sw0x/uy**" (con x=1,2,...,8 e y=0,1). Usando il nome "sw0x" il file system viene montato via HPS, altrimenti via Ethernet. Gli stessi pathname sono attivi sulla CW, ma il mount viene sempre effettuato via Ethernet.

Vulcan

- Il file system **“/u0” di sp01** contiene le home directories degli utenti
- Su tutti gli altri file system gli utenti possono allocare spazio disco, creando una propria directory (si suggerisce di usare come nome il proprio userid)
- I nodi e la CW appartengono ad un unico dominio NIS (note anche come YP: Yellow Pages)
- Gli stessi userid e passwords sono disponibili su tutto il dominio. La password puo’ essere modificata con il comando “**yppasswd**”

Vulcan

- I singoli nodi possono essere raggiunti mediante un **telnet** all'indirizzo "**sp0x.hpc.unipg.it**" (con x=1,2,...,8), ma se ne sconsiglia l'uso interattivo.

- Alla CW si puo' accedere con un telnet a "**vesta.unipg.it**" o a «**vulcan.unipg.it**».

- Tutto il lavoro interattivo (editing di files, compilazioni, debugging, sottomissione di jobs, etc.) va preferenzialmente svolto sulla Control Workstation.

- La sottomissione di jobs (batch) sui nodi e' controllata dal **LoadLeveler**. L'uso interattivo dei nodi e' permesso solo per piccoli job di breve durata (a scopo di debugging).

Sistema Operativo

- Il software di base dell'SP2 e' AIX/6000 con tutti i suoi componenti: Sistema, TCP/IP, NFS, Compilers, etc.

- I nodi SP2 girano codice eseguibile RISC/6000 e sono quindi compatibili a livello binario con gli eseguibili RISC/6000

- Un nodo SP2 puo' esser coinvolto nell'esecuzione di job **paralleli** o eseguire un "piu' tradizionale" job **"seriale"**: da questo punto di vista l'SP2 puo' essere considerato equivalente ad un "cluster" di RISC/6000 connesso con una rete ad alta velocita'

- L'ambiente di lavoro ed i comandi di base sono quelli di una normale WS **UNIX** (AIX)

Parallel System Support Program

Il “Parallel System Support Program”, **PSSP** e’ una collezione di programmi che consente di amministrare l’SP2.

Contiene tutto il software necessario per installare, gestire e mantenere il sistema SP2 dalla Control Workstation.

Del PSSP fanno parte alcuni packages di pubblico dominio: amd, sup, perl, etc.

Consente il monitoring costante della situazione dei nodi e dello switch.

Uno dei suoi componenti e’ il “**Resource Manager**”: si occupa di allocare “pools” di nodi che soddisfino le richieste delle applicazioni (per esempio nel caso dell’HPS «user space mode» di individuare un nodo libero).

Parallel Environment

- Il “Parallel Environment” contiene i componenti software necessari per:
  - sviluppare

    MPL “**Message Passing Library**” e’ una libreria applicativa (una API: Application Programming Interface). Nei nuovi release software e’ presente anche il nuovo “standard” **MPI “Message Passing Interface”**
  - eseguire

    Parallel Operating Environment (**POE**)

    permette di eseguire job paralleli interagendo con il Resource Manager per l’allocazione dei nodi
  - analizzare le perfomance

    Visualization and Perfomance Monitoring Tool (**VT**)

Parallel Environment

- eseguire il debugging

  Parallel debuggers

  pdbx estende le funzionalita' di dbx ad applicazioni parallele

  xpdbx fornisce una interfaccia grafica X

  di "**Applicazioni Parallele**".

- Fornisce le funzioni necessarie ad un *ambiente applicativo parallelo*.

- Le applicazioni MPL (in futuro MPI) possono comunicare via Ethernet, via HPS IP mode e via HPS user space mode.

- Le librerie MPL,MPI possono essere chiamate da programmi scritti in linguaggio Fortran o C.

PVM3

- Il PVM3 (Parallel Virtual Machine) e' un package di pubblico dominio distribuito dall'Oak Ridge National Laboratory

- E' una delle librerie "message passing" (Fortran ,C) piu' diffuse su cluster di workstations e su macchine parallele a memoria distribuita.

- Su Vulcan e' installata la versione 3.3.7 (con **PVM_ROOT** in **/usr/sp/pvm3**)

- Il PVM3 standard puo' comunicare via Ethernet e unicamente via HPS IP mode (***scegliendo i nomi sp0x o sw0x***). Le applicazioni PVM3 **non possono** utilizzare l'HPS in user space mode.

PVMe

- Il prodotto IBM **PVMe AIX** e' l'implementazione IBM del PVM3 ed e' compatibile con il PVM 3.3 di Oak Ridge.

- Le applicazioni devono usare le librerie del PVMe invece che quelle del PVM3.

- Il PVMe usa lo switch in "user space mode" e fornisce alle applicazioni PVM prestazioni ottimali dell'HPS.

- Dalla versione 3.3.8 il PVM3 puo' appoggiarsi sulla libreria **MPI**, disponibile nella *versione 2 del Parallel Environment*, fornendo quindi funzioni equivalenti al PVMe.

LoadLeveler

Il LoadLeveler e’ un job-scheduler distribuito.

Permette di bilanciare il carico di lavoro sui nodi di un SP2 e/o di un cluster di workstations.

Gestisce lo scheduling sia di job seriali che paralleli ed e’ in grado di interagire con il Resource Manager per cio’ che riguarda l’allocazione dei nodi.

LoadLeveler permette di:

- Sottomettere e cancellare jobs
- Monitorare lo stato dei job
- Modificare le priorita’ dei job

con comandi UNIX e per mezzo di una interfaccia grafica X.

LoadLeveler

Su vulcan i comandi del LoadLeveler possono essere eseguiti dal path “/**usr/sp/loadl/bin**” o dal path «**/usr/lpp/LoadL/nfs/bin**»:

- **llsubmit**

  sottomissione di jobs

- **llq**

  monitoring dello stato dei job

- **llstatus**

  monitoring dello stato dei nodi

- **llcancel**

  cancellazione di jobs

- **xloadl**

  interfaccia grafica X

LoadLeveler su Vulcan

- I job sottomessi al LoadLeveler vengono raggruppati in “**classi**”
- Attualmente al LoadLeveler di Vulcan (provvisorio) sono definite tre classi di jobs

  small jobs fino a 50Mb

  medium jobs fino ad 100Mb

  large jobs fino a 200Mb

  e su ogni nodo puo’ girare un solo job per ciascuna classe.
- La sottomissione di job al LoadLeveler avviene mediante la preparazione di opportune scripts che contegono comandi LL. Esempi possono essere trovati in “**/usr/sp/loadl/samples**”.
- La guida utente completa del LL (in formato Postscript) puo’ essere stampata da “/**usr/lpp/LoadL/postscript/lluser.ps.Z**” sulla CW.

Compilatori e librerie applicative

- Su vulcan sono installati i compilatori:
  - IBM XLF Fortran 3.2.2 (9 licenze)
  - IBM XLC 1.3.0
- e le librerie applicative
  - IBM ESSL 2.2.1
  - NAG
- Sono gli stessi prodotti software presenti sulle workstations RISC/6000 e possono essere usati esattamente nello stesso modo

Documentazione

- I manuali dei vari componenti software sono accessibili direttamente sulla Control Workstation (sui nodi non e' presente «tutta» la documentazione)

- L'applicazione Info Explorer permette di effettuare ricerche e di ottenere copie delle pagine dei manuali. Si lancia con il comando «**info**». Se ne suggerisce l'uso in ambiente X/Motif.

- Il tradizionale comando UNIX «man» fornisce l'accesso alle pagine dei manuali dei vari comandi.

Documentazione

- La variabile di environment MANPATH consente di definire il sottoinsieme di pagine a cui si vuole far riferimento (usa la stessa sintassi della variabile PATH).

  Alcuni MANPATH utili:

  - **/usr/man** oppure **/usr/share/man**

    Sistema Operativo AIX, Par. Env, PVMe

  - **/usr/lpp/ssp/man**

    Parallel System Support Program

  - /usr/lpp/LoadL/nfs/man oppure /usr/sp/loadl/man

    LoadLeveler

  - **/usr/sp/pvm3/man**

    Oak Ridge PVM 3.3.7

Documentazione

- Le copie dei manuali in formato postscript sono disponibili in (file .ps o .ps.Z):
  - /usr/lpp/ssp/docs

    Parallel System Support Program
  - /usr/lpp/pedocs

    Parallel Environment: MPL, POE, VT, PDBX
  - /usr/lpp/pvm3/pvmeug.ps

    Guida utente PVMe
  - /usr/sp/pvm3/postscript

    PVM3
  - /usr/lpp/LoadL/postscript

    LoadlLeveler

  nella Control Workstation «vesta.unipg.it».

PATH utili su vulcan

- Oltre ai PATH standard su una macchina AIX questi sono alcuni PATH utili su «vulcan»:

/usr/lpp/ssp/bin	Parall. Supp. Program
/usr/lpp/poe/samples	Esempi Parall. Env.
/usr/lpp/pvm3	PVMe
/usr/sp/pvm3	PVM3, XPVM
usr/sp/loadl/bin	LoadLeveler
/usr/lpp/LoadL/nfs/bin	
/usr/local/bin	ambiente locale AIX
/usr/sp/bin	ambiente locale SP2

- I comandi del Parallel Environment sono installati sui PATH standard AIX (/bin, /usr/bin, ...).

Alcuni comandi utili

- Alcuni comandi utili su «vulcan»:

spsystat	stato dei nodi
sptop	stato dei processi
(package di pubblico dominio aixclmon)	
nodes_up	nodi funzionanti
pvme	PVM

  sono disponibili nel PATH /usr/sp/bin.

Informazioni via WEB

- Alcune fonti di informazioni interessanti possono essere raggiunte sul WEB, con un WWW browser (Netscape, Mosaic, etc.):
  - **International Business Machines**

    http://www.ibm.com/
  - **IBM High-Performance Computing**

    http://ibm.tc.cornell.edu/index.html
  - **IBM Planetwide search**

    http://www.austin.ibm.com/Search/
  - **IBM RISC System/6000**

    http://www.rs6000.ibm.com/
  - **Maui High Perfomance Computing Center**

    http://www.mhpcc.edu