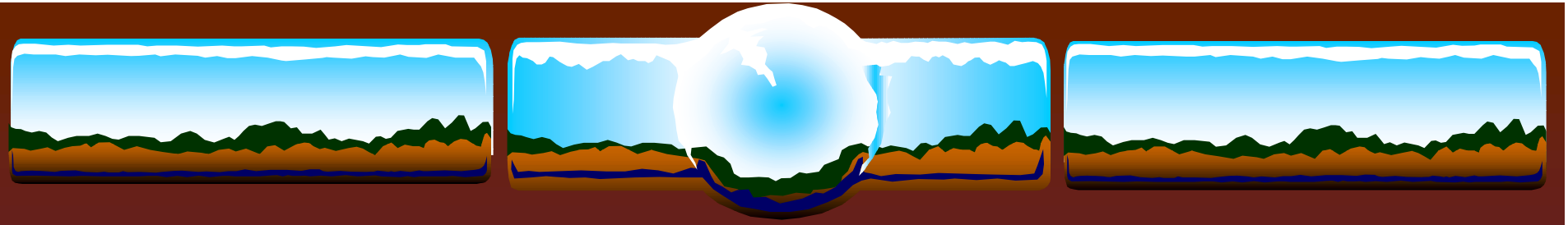


Beowulf Cluster Administration on Linux and HP-UX

Roger Goff
Solutions Specialist

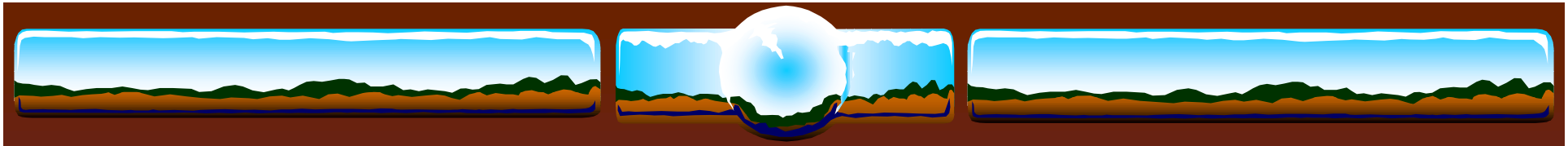
Hewlett-Packard Company
America's Technical Computing
Advanced Technology Center

Roger_Goff@hp.com
3404 East Harmony Road
Fort Collins, CO 80528
(970)898-4719 FAX (603)590-5337

Dr. Randy Splinter
Solutions Specialist

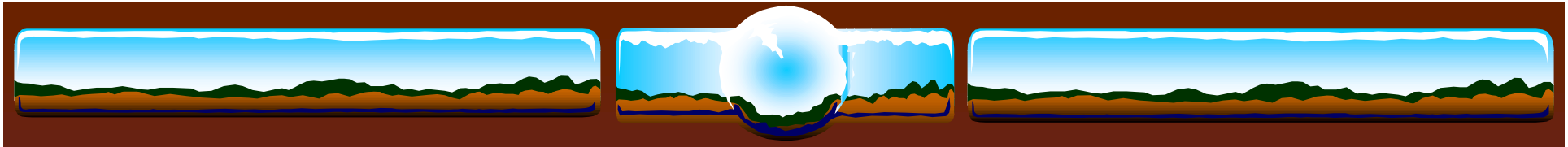
Hewlett-Packard Company
America's Technical Computing
Advanced Technology Center

Randy_Splinter@hp.com
238 Serenoa Drive
Canton, GA 30114
(404)648-8003 FAX (678)493-8103



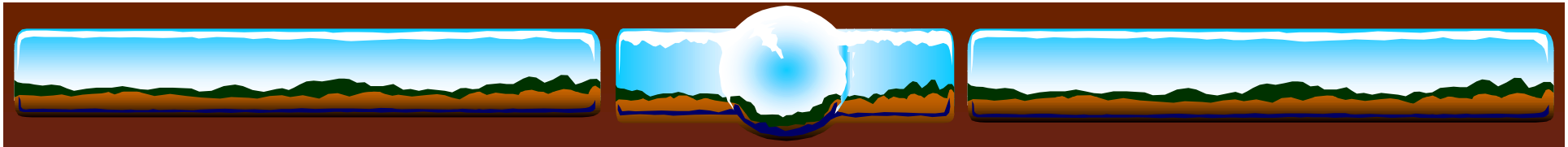
Topics to be covered:

- Beowulf Cluster Definition
- The physical layer – how to connect to the compute nodes in a cluster, including a cost analysis of the approaches
- The logical layer – how to administer and monitor a cluster



What is a Beowulf Cluster?

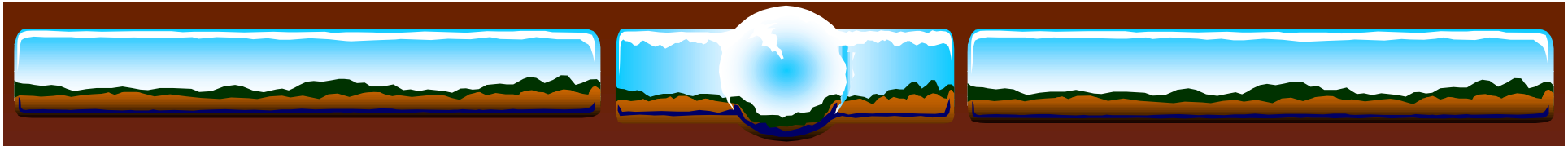
A kind of high performance, massively parallel computer, interconnected by a private, high-speed network. It consists of a cluster of PCs or workstations dedicated to running high-performance computing tasks. The nodes in the cluster don't sit on people's desks; they are dedicated to running cluster jobs. It is usually connected to the outside world through only a single node.



Beowulf Cluster vs. N.O.W.

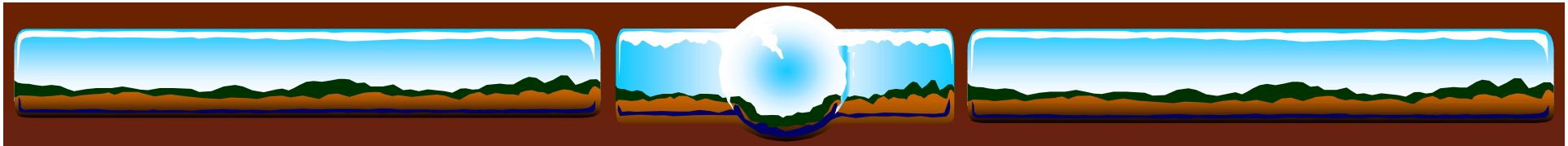
A Beowulf cluster is distinguished from a *Network of Workstations* (N.O.W.) by several subtle but significant characteristics:

- The nodes in the cluster are dedicated to the cluster
- The cluster node interconnect is isolated from the external network, thus network load is determined by the application being run.
- Clusters do not have authentication concerns between the processors in the system
- OS parameters may be tuned to improve application performance without concern for impacts on general system performance



Beowulf Clusters

- For more information
  - http://www.beowulf.org
  - http://www.beowulf-underground.org

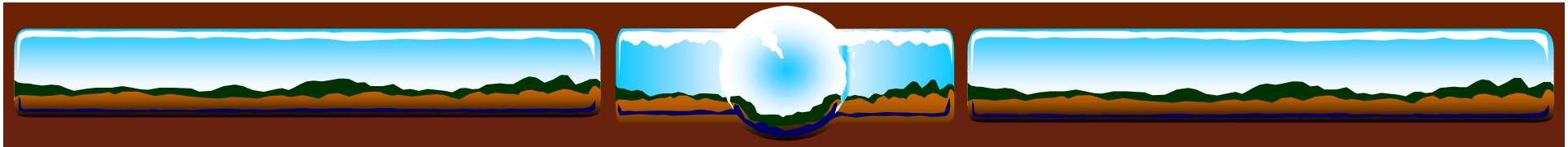


Beowulf Cluster Administration - The Physical Layer

For the purposes of this discussion we are only considering accessing the individual compute nodes in the cluster beyond logging in to a node over the network. One may argue that if you can't telnet into a machine over the network you may as well physically recycle the power anyway, but it may not always be possible or desirable to take this approach.

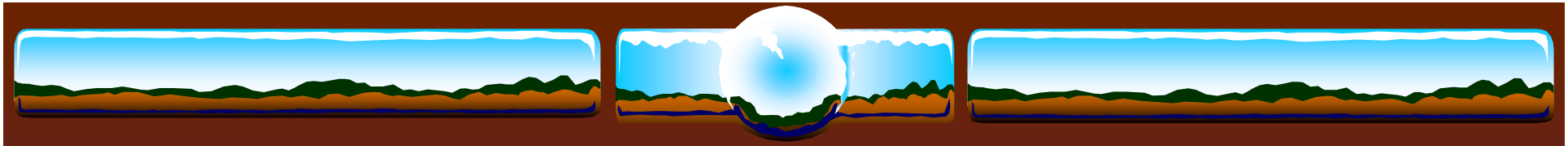
Approaches to consider:

- Local, in computer room, access
- Remote access (beyond telnet)



Local Access

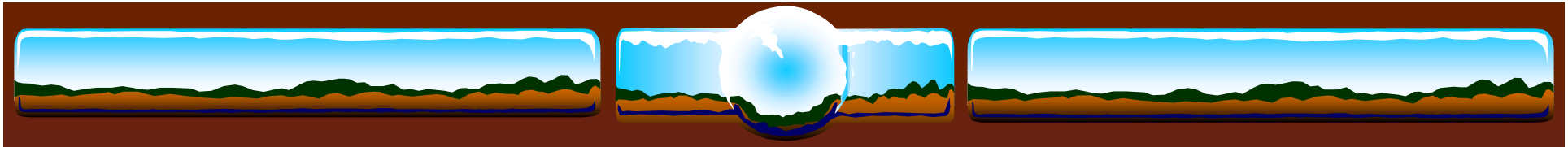
- Use Keyboard, Video, Mouse (KVM) switch
  - Pros:
    - Up to 256 nodes can be administered from one console
    - Less expensive than providing remote administration capabilities for small numbers of nodes (less than ~40)
  - Cons:
    - Must administer each system from within computer room
    - Cannot administer nodes over the network if network access to a node goes away
    - Can’t access IPMI capabilities (more on this later)
    - Clusters limited to 256 node (without adding a second keyboard, monitor and mouse station for administration)



Remote Access

Uses an *administrative node* along with a terminal server

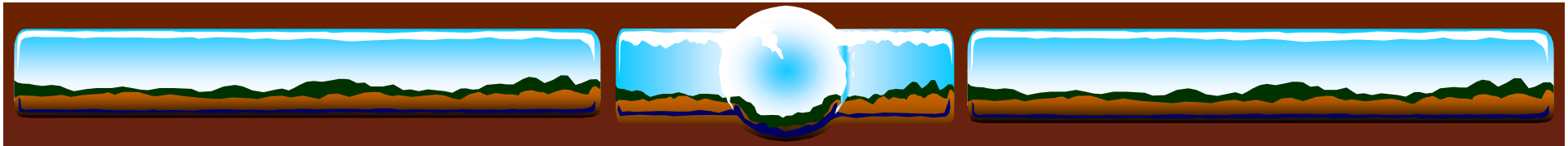
- Pros:
  - Can administer nodes from anywhere on the network even if the cluster or node's network is down (you must be able to reach the administrative node over the network)
  - One administrative node can access any number of cluster nodes
  - An administrative node can serve other purposes like file serving, installation services, cluster job management and running cluster management software
  - Can access IPMI capabilities (more on this later)
- Cons
  - Expensive solution for a small number of nodes (less than ~40) unless you can use an existing machine in your environment for the tasks
  - System access speed limited to speed of serial terminal port



Intelligent Platform Management Interface

- IPMI was co-developed by Intel and HP to provide a low level consistent management interface for hardware vendors to provide with their systems
- IPMI support is built in to certain Intel motherboards which are used by many platform vendors today
- HP's NetServer products have IPMI support built in
- IPMI provides for the ability to remotely control:

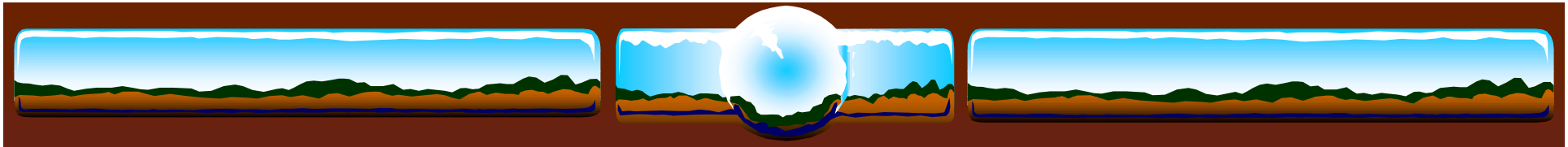
  Device power control, hardware reset, remote bios console, power system monitoring, chassis intrusion monitoring, fan status monitoring and more.
- IPMI is implemented in hardware running on separate micro controllers, so monitoring does not impact a node's performance



Cost Comparison

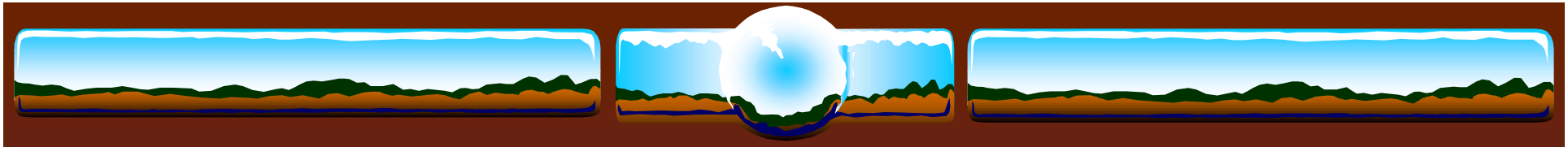
	Local Access Solution Cost	Remote Access Solution Cost
16 nodes	$ 900-1600 KVM switch $ 600 cables $1500-2200 total	$3500 admin node $ 144 Cat5 cable + connectors $1400 32 port terminal server $ 5,044 total
32 nodes	$1800-3200 KVM switch(s) $1260 cables $3060-4460 total	$3500 admin node $ 336 Cat5 cable + connectors $1400 32-port terminal server $ 5,236 total
48 nodes	$5400-9600 KVM switch(s) $2200 cables $7600-11800 total	$3500 admin node $ 544 Cat5 cable + connectors $2800 2x32-port terminal servers $ 6,844 total
64 nodes	$6000 KVM switch(s) $4200 cables $2500 KVM monitor $12,700 total	$3500 admin node $ 760 Cat5 cable + connectors $2800 2x32-port terminal servers $7,060 total

Assumption: racks of 2U compute nodes (20 per rack)



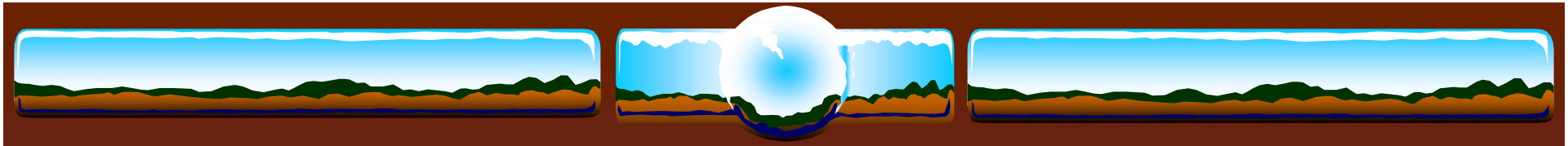
Access Cost Analysis - Local vs. Remote

- Cost of cabling is the key driver:
  - KVM cabling costs
    - 6 foot - $35
    - 10 foot - $45
    - 15 foot - $55
    - 25 foot - $75
  - Terminal servers use CAT5 or Telco cabling which is cheap
- Density of racks in computer room becomes a factor
- Providing remote access will be cheaper than local access for large clusters and when you already have a system that can perform the required administrative tasks



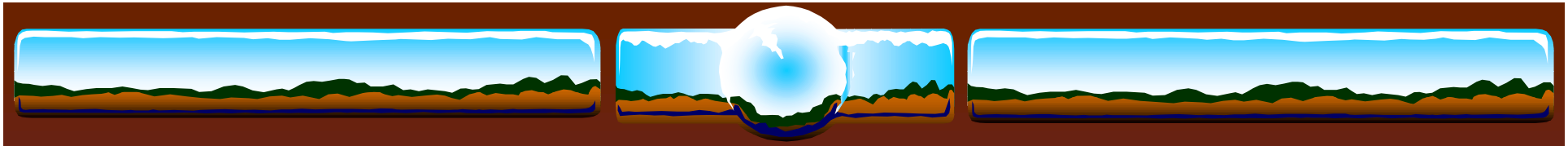
The Logical Layer - golden images

To ensure consistent results in computation and performance it is necessary that the software for each compute node in a cluster be configured identically. This is typically achieved by configuring and tuning one system to a known good state and then replicating that state to all other nodes in the cluster. We call the software clone of a known good state a “golden image.”



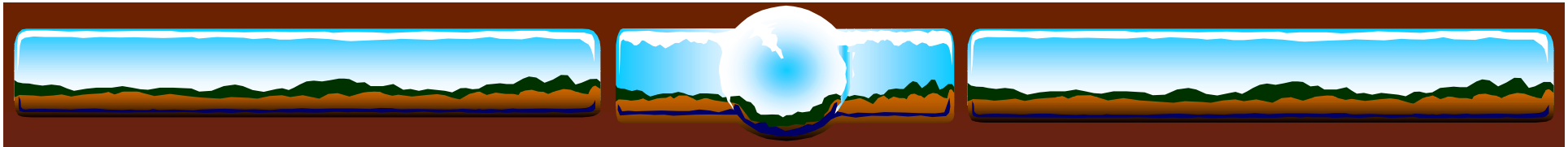
Golden Images

The process for creating the actual image to be replicated differs from HP-UX to Linux in the tools that are available. The tool capabilities and modes of operation differ as well. In the next few slides we will look at HP-UX's Ignite-UX and the Linux tools *Kickstart* from Red Hat® and *SystemImager* from VA Linux Systems®



Ignite-UX for HP-UX

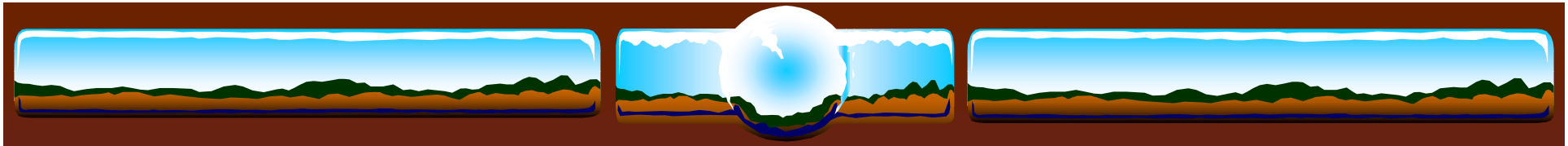
- Built by HP for HP-UX
- Can do:
  - Attended installs via a GUI
  - Unattended package/depot based installs locally or across the net
  - Golden image installs from CD or across the net
  - Dynamic per client customizations based on client hardware
  - *bootsys* can take control of a client for either automatic or unattended installations
- Ignite-UX drawbacks:
  - Currently only runs on HP-UX



Ignite-UX for HP-UX

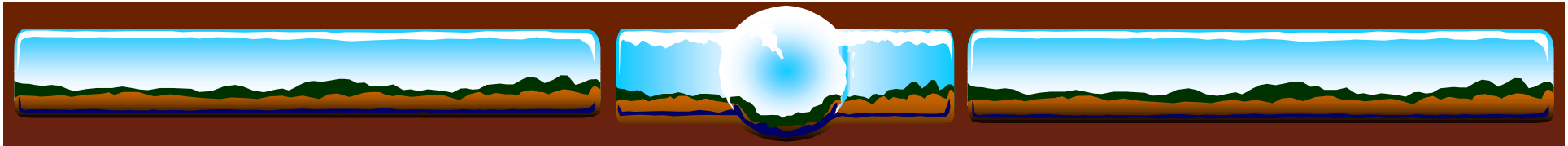
- Image Creation Process
  - Configure and test a client to use as your golden image source
  - *make_sys_image* creates a golden image of the client that includes all customizations
  - *archive_impact* – creates volume size information for the file system volume groups of your golden image needed in the golden image configuration file on your Ignite-UX server
  - Customize the Ignite-UX server configuration file for your golden image
- For more information:

  http://www.software.hp.com/products/IUX/docs.html



Kickstart for Linux

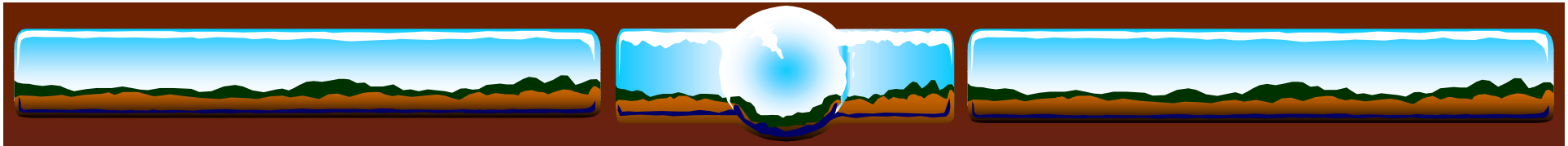
- Built by and for Red Hat® Linux
- Provides unattended , across network installations, capable of repartitioning disk drives
- Essentially a scripted Red Hat® Linux install with post installation customization
- Red Hat® kickstart drawbacks:
  - Can’t do true “golden image” based installs
  - Post install customization scripts can be difficult to debug
  - No other Linux distributions supported
  - No GUI



Kickstart for Linux

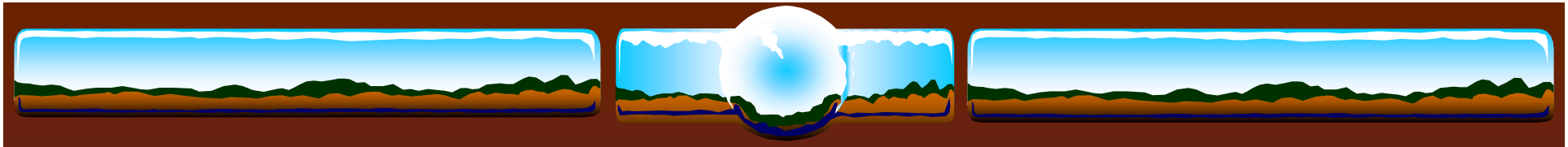
- Image Creation Process
  - Configure and test a client to use as your golden image source
  - *mkkickstart* – creates a kickstart configuration file that includes all configuration information for the golden image
  - Create post install customizations in kickstart configuration file
  - Create boot media from *boot.img* or *bootnet.img* images that you customize to invoke kickstart
- For more information:

  http://redhat.com/support/manuals/



SystemImager for Linux

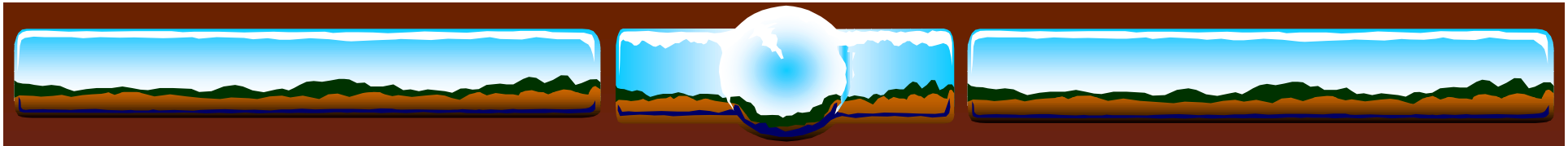
- Built by VA Linux Systems
- Linux Distribution independent
- Provides unattended, across network installations, capable of repartitioning disk drives
- Golden image installs (not a package loader)
- SystemImager drawbacks:
  - Can’t do package/depot based installs
  - No GUI



SystemImager for Linux

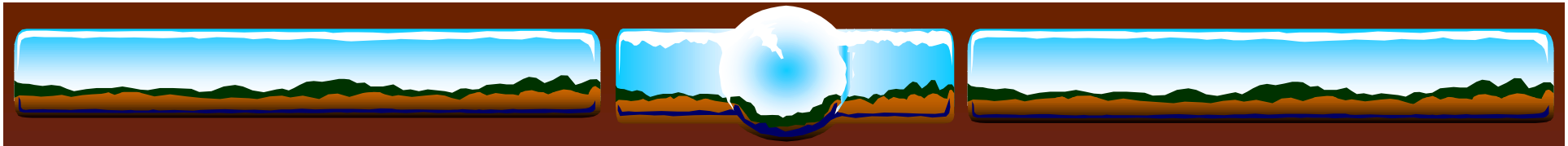
- Image creation process:
  - Configure and test a “master client” system
  - *preparecclient* – run on master client collects disk partition information and configures *rsync* server on client system
  - *getimage* – run on “image server” to pull files from master client and create an auto installation script. Can also be used to “update” your golden image after changes are made to your master client
  - *makeautoinstallcd* and *makeautoinstalldiskette* – build boot media customized to access image server
- For more information

  http://systemimager.org



Monitoring Tools

- Why monitoring tools?
  - Traditional methods of monitoring single systems will not work effectively
- Network Health Tools
- Node Health Tools
- Resource Monitoring Tools
- Caveats
  - I do not intend this to be an exhaustive overview
  - My goal is to cover enough of a variety of tools to give some flavor what is available

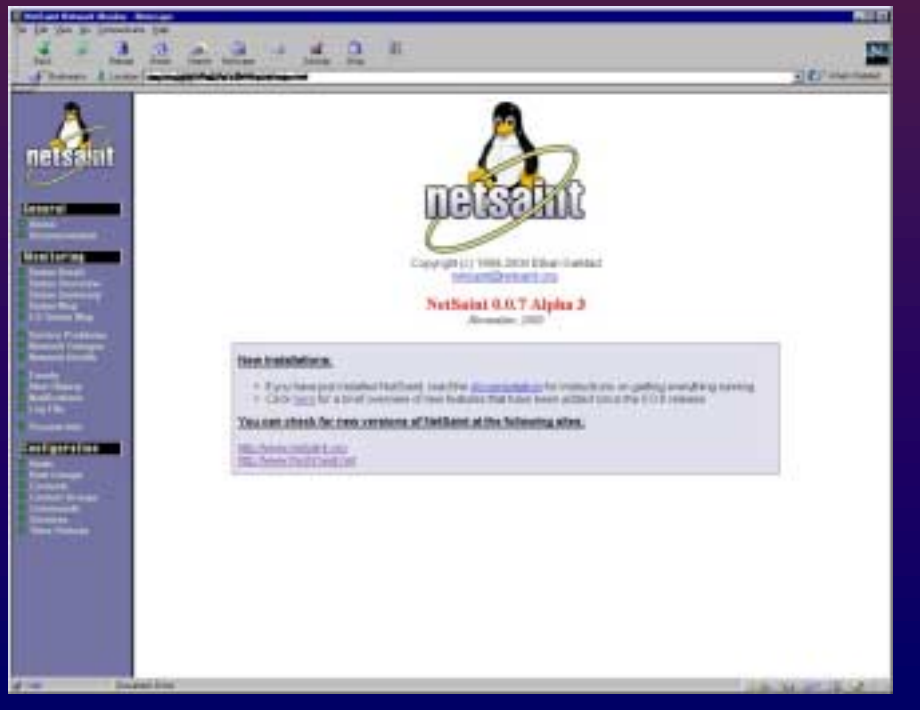


Network Health and Monitoring Tools

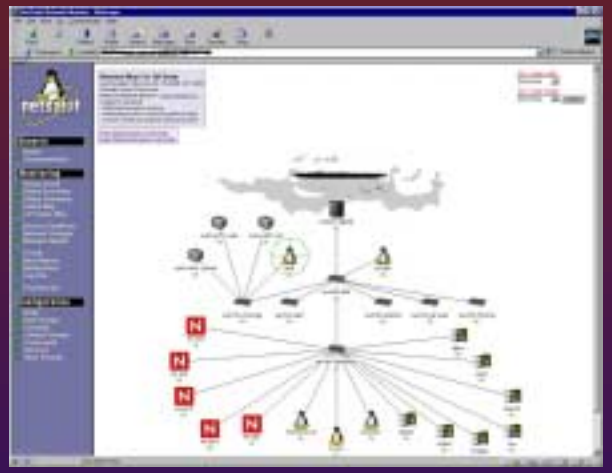
- NetSaint (http://www.netsaint.org)
- Big Brother (http://bb4.com)
- Ntop (http://www.ntop.org) or RedHat distribution CDs
- HP Network Node Manager

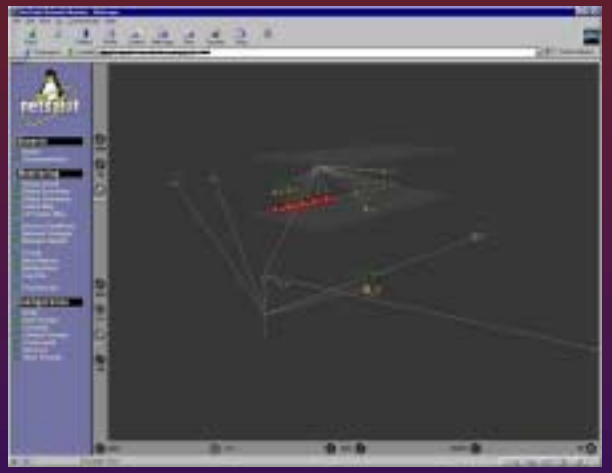
Net Saint

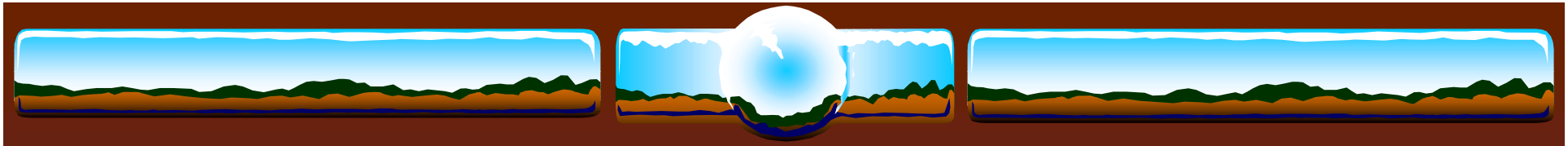
- Net Saint is a freeware tool downloadable from http://www.netsaint.org
  - Monitors hosts and services including disk, memory usage, processes, log files, etc
  - NetSaint can email or page when a problem shows up
  - Can automatically run scripts when a problem shows
  - Web interface available
  - Supports monitoring hosts



Net Saint Screenshots





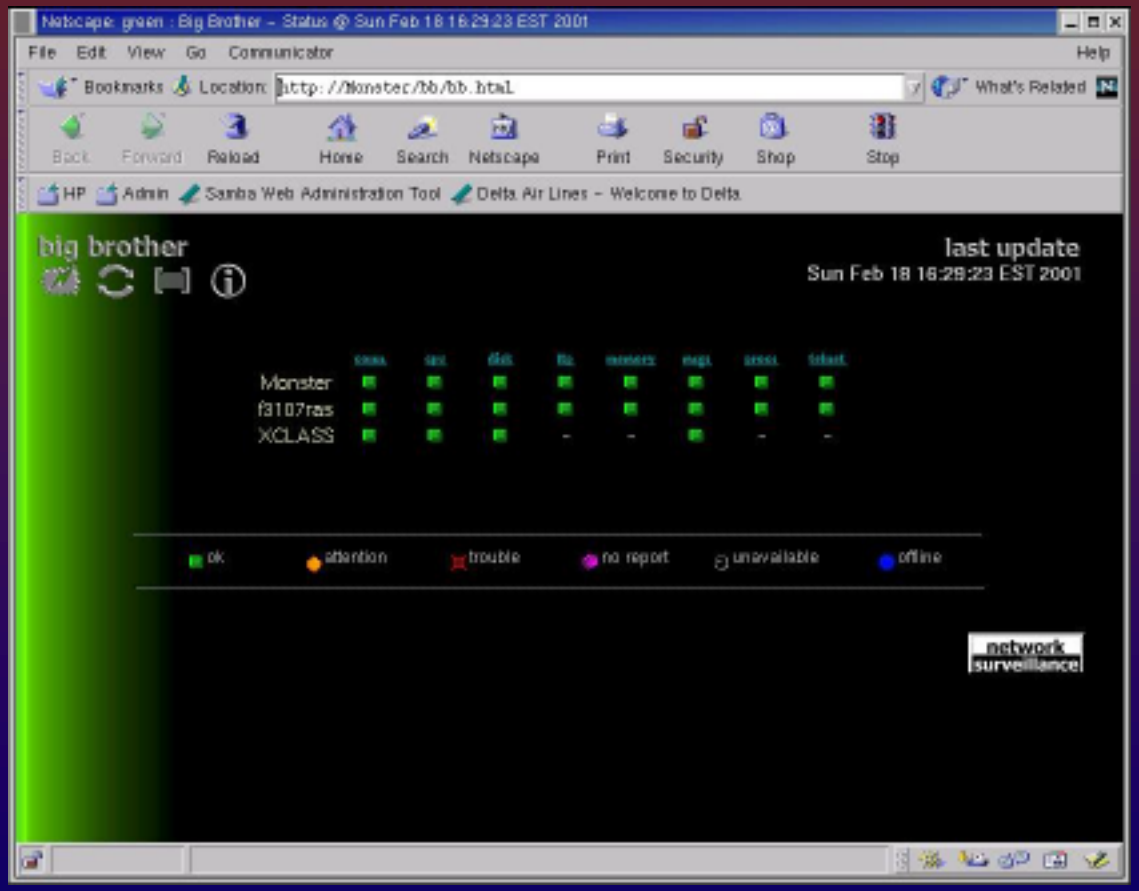


Big Brother

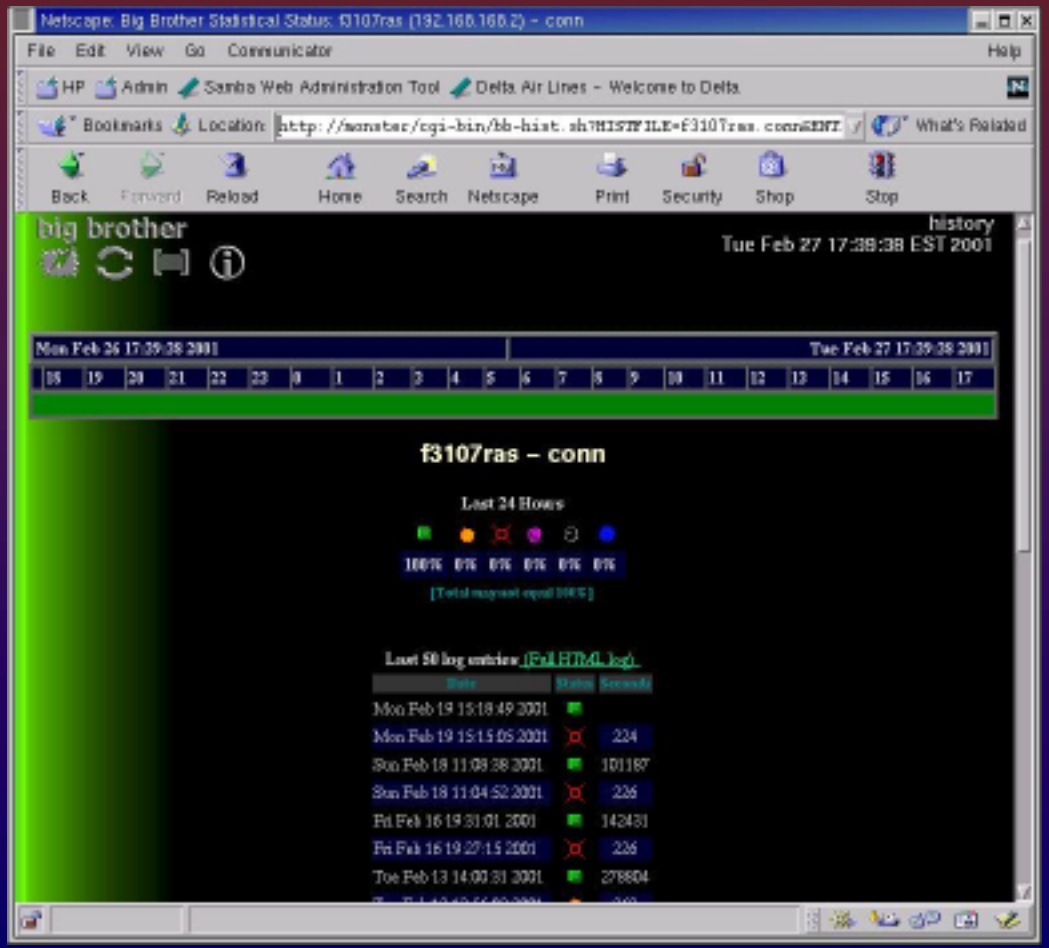
- BB is a freeware tool downloadable from http://www.bb4.com
- BB is a web-based monitoring tool
  - It requires the server run on a system that also has a web server running
- There is currently a
  - UNIX server and client
  - Linux server and client
  - NT client (not fully functional, not surprising?)
- The UNIX server and client compiled without problems on both HP-UX and RedHat 7.0
- BB is not very sophisticated, but gives you a simple quick view on the overall health of the cluster

Big Brother Screen Shots

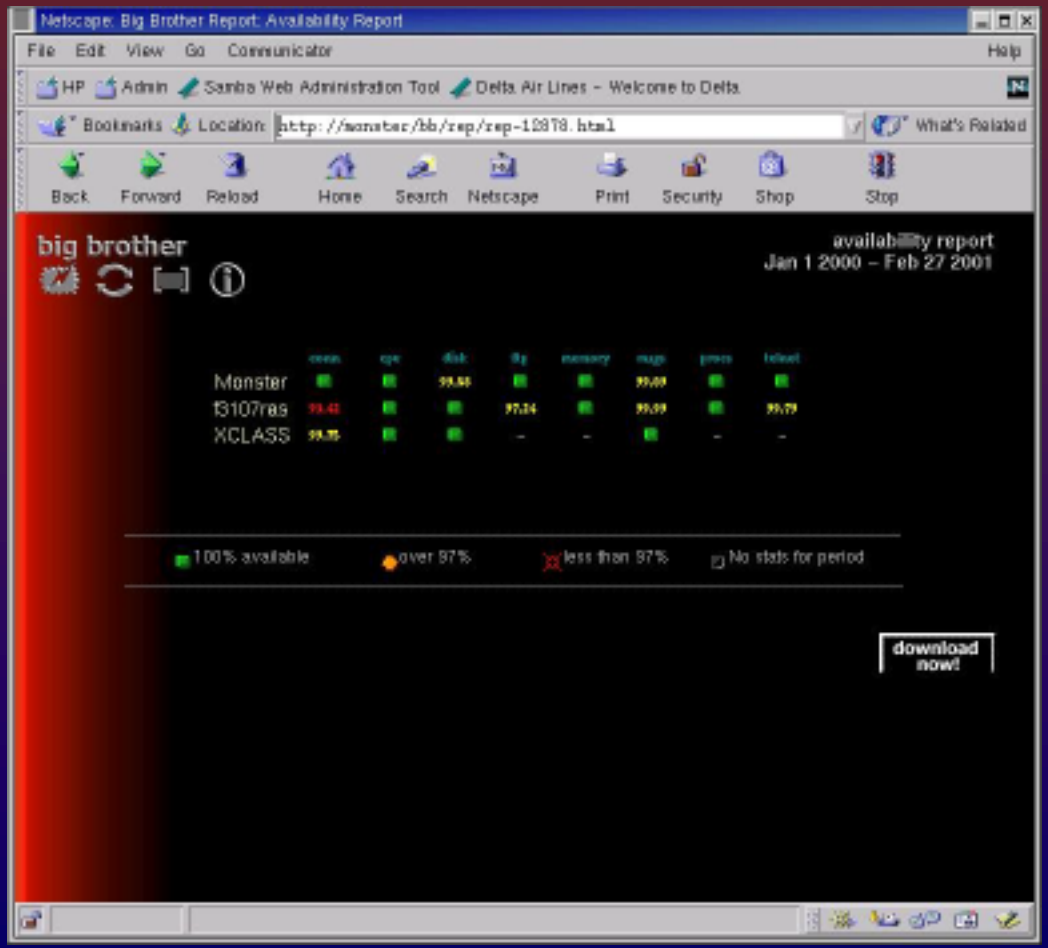
- Main Screen



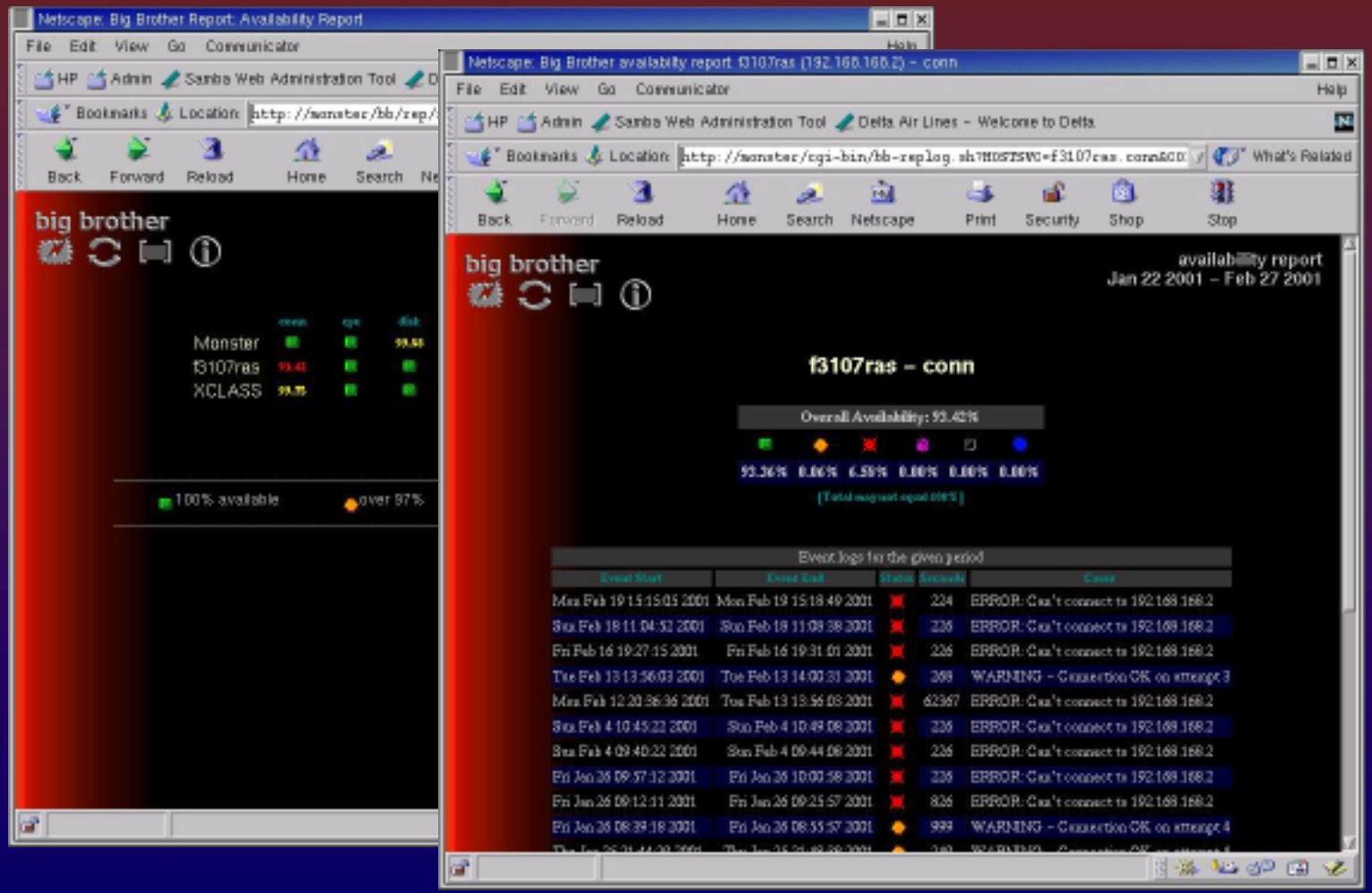
Network Connectivity History (past 24 hours)



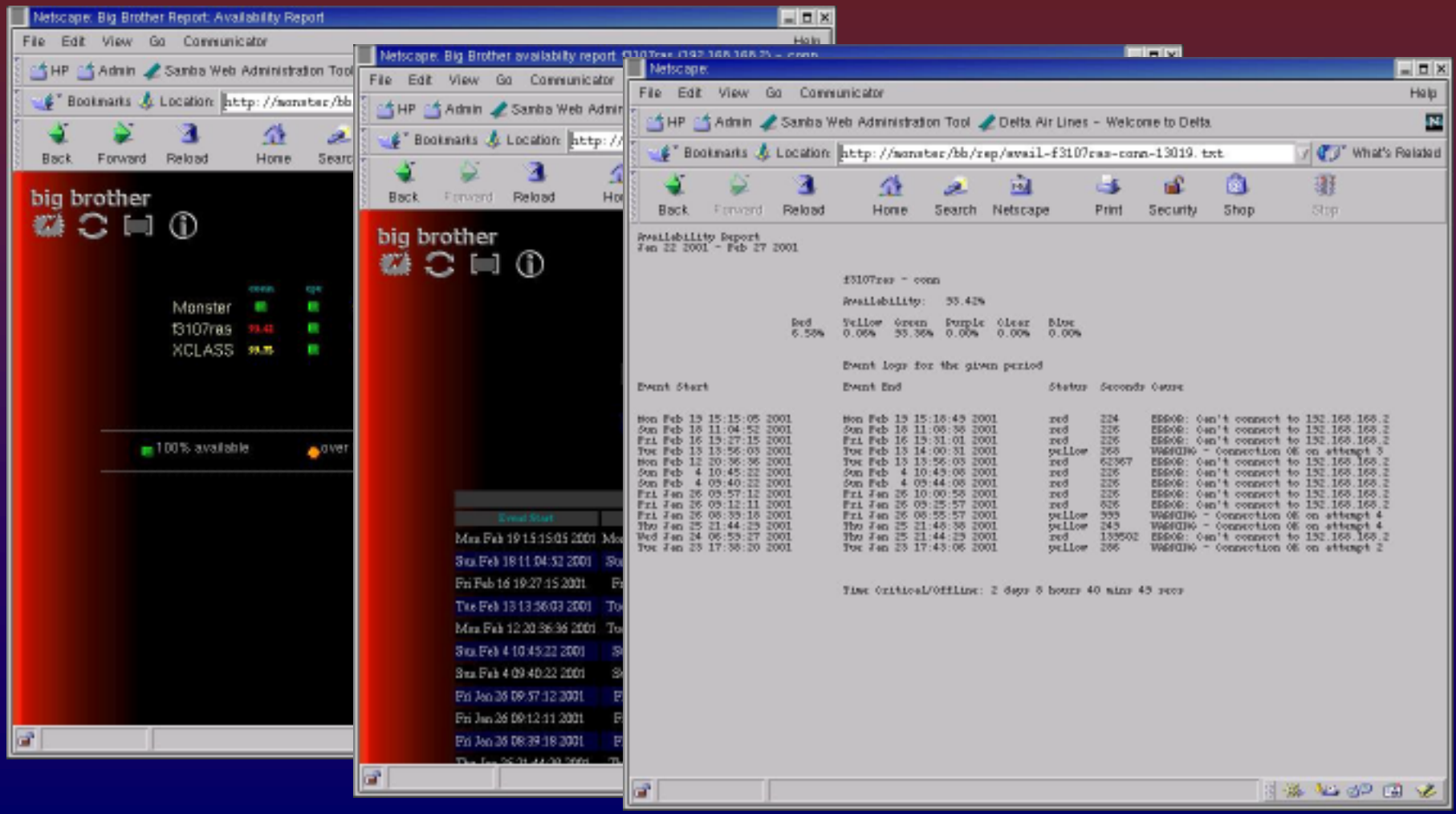
Network Connectivity Report



Network Connectivity Report (Details)



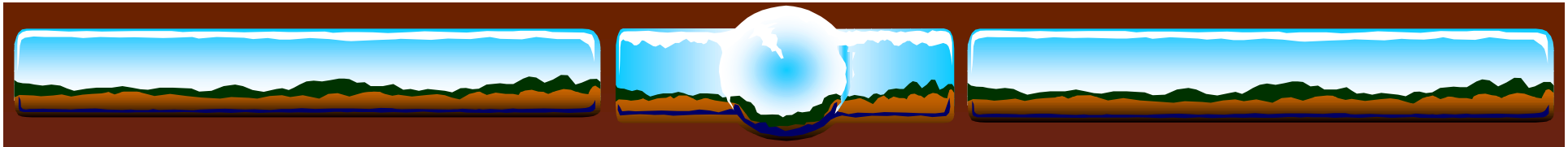
Network Connectivity Report (text format)



Ntop

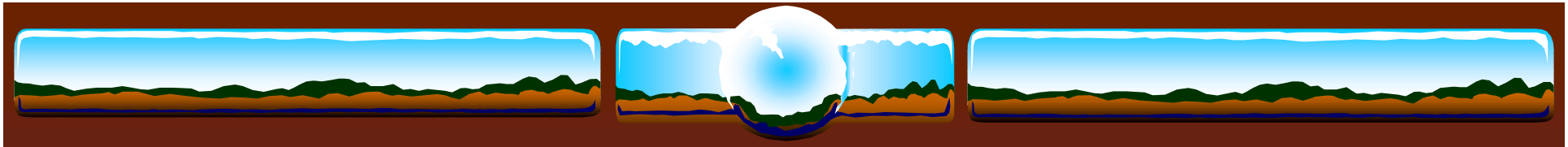
- Ntop is a tool for displaying network usage
  - Has both a curses interface and a web interface
- Freely downloadable from http://www.ntop.org


```
                intop 0.0.1 (May 19 2000) listening on [hme0]
6606 Pkts/770.7 Kb [IP 703.7 Kb/Other 67.1 Kb]     Thpt: 211.9 Kbps/349.7 Kbp
Host            Act   -Rcvd-      Sent        TCP        UDP       ICMP
more            B   257.4 Kb   281.9 Kb   256.6 Kb        769          0
zetant          B   204.2 Kb   232.3 Kb   204.2 Kb          0          0
tar             B    42.9 Kb    19.5 Kb    42.9 Kb          0          0
ibook           B    32.7 Kb     4.7 Kb    32.7 Kb          0          0
tecserv         R        791          0          0        595        196
bugnoli         B        602     1.4 Kb          0        602          0
urano           B        496     5.1 Kb          0        496          0
utlrouter       R         98          0          0          0         98
mis             S          0        212          0          0          0
fiorella        S          0        486          0          0          0
piutltst02      S          0     1.4 Kb          0          0          0
mostardi        S          0        952          0          0          0
193.43.104.55   S          0        588          0          0          0
itest1          S          0        928          0          0          0
rolly           S          0         46          0          0          0
itin2           S          0         92          0          0          0
3comhub1        S          0        610          0          0          0
re              S          0     5.6 Kb          0          0          0
pi100           S          0     1.2 Kb          0          0          0
lcardini        S          0        546          0          0          0
mbeng           S          0        602          0          0          0
itest2          S          0        600          0          0          0
fossati-a       S          0        960          0          0          0
hpwsutl         S          0     3.1 Kb          0          0          0
catlc           S          0        120          0          0          0
aut01b          S          0        243          0          0          0
biu             S          0        542          0          0          0
artico2         S          0        226          0          0          0
```



HP OpenView Network Node Manager

- NNM is part of the HP OpenView suite of products for enterprise management
- Supplies much of the same functionality as both Net Saint and Big Brother
- Probably not a good choice unless
  - The cluster is to be integrated into the enterprise environment
  - A well supported tool is required
  - The tool needs to be integrated into either OpenView or another enterprise management suite, such as Tivoli

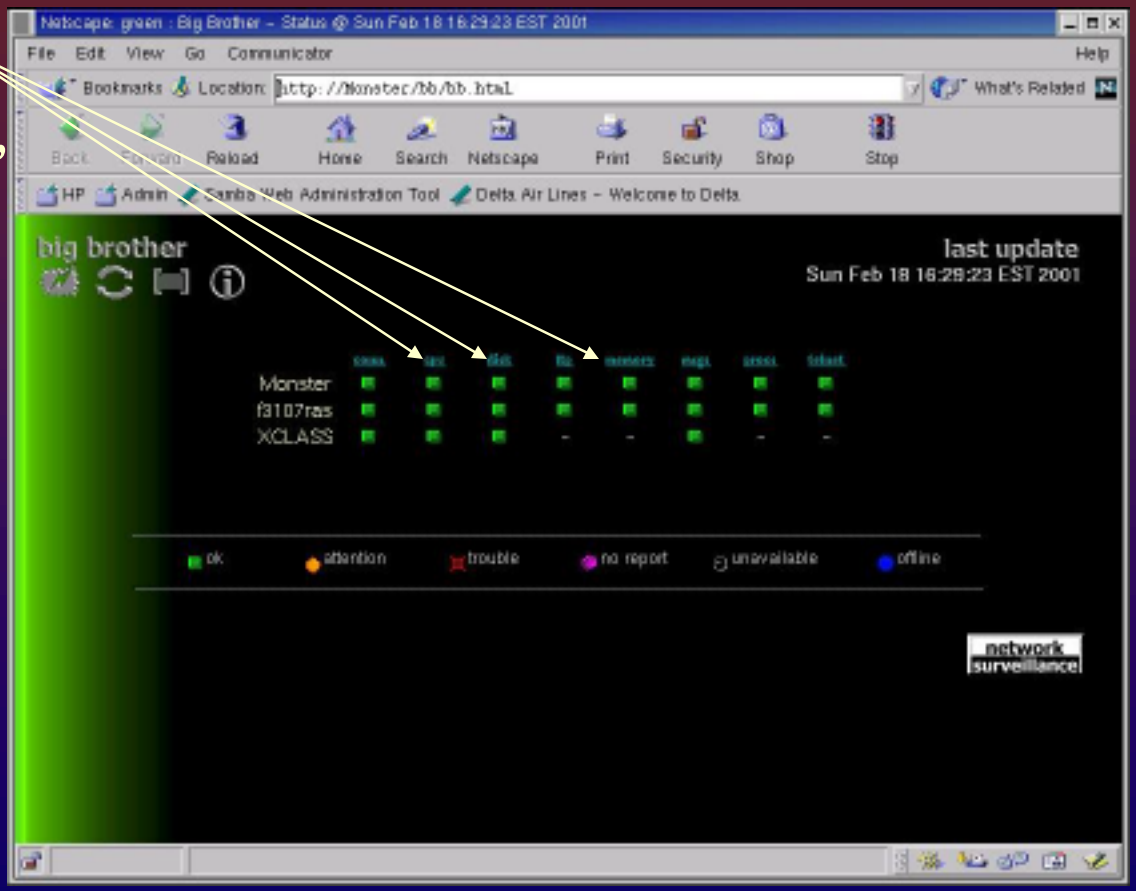


Node Health Tools

- Net Saint (http://www.netsaint.org)
- Big Brother (http://www.bb4.com)
- Xosview (http://lore.ece.utexas.edu/~bgrayson/xosview.html) or RedHat distribution CDs
- HP Glance

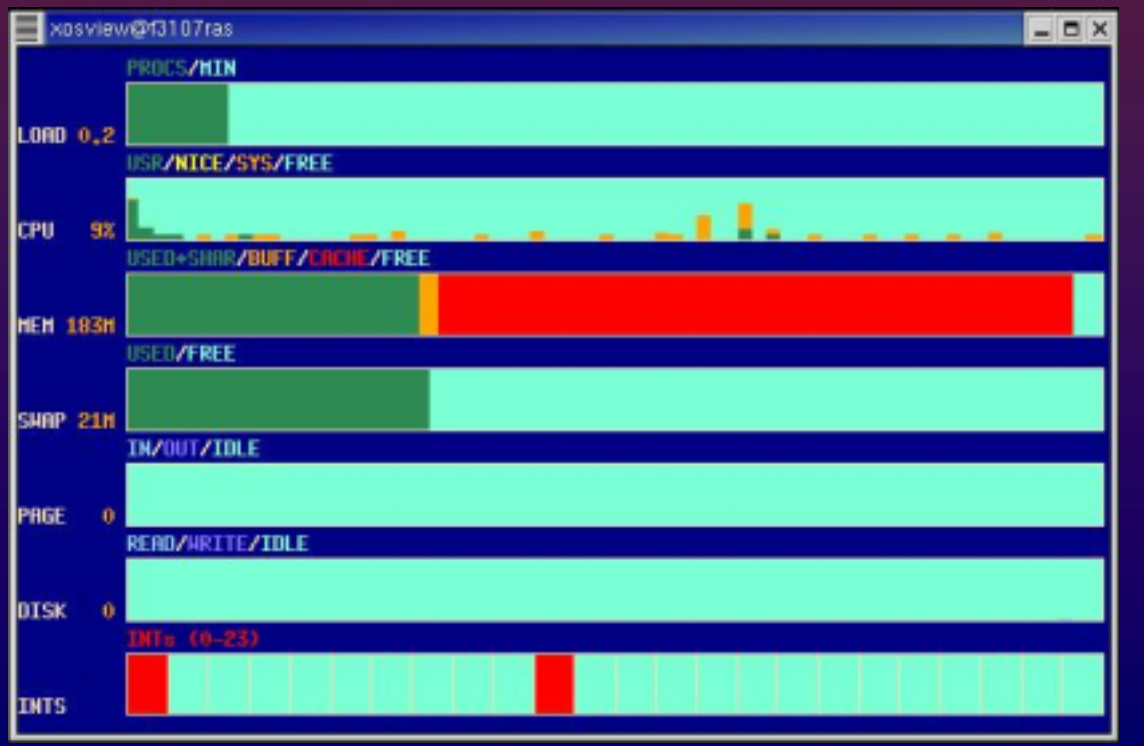
Big Brother Node Health

- Information on the CPU usage, disk usage, memory usage all can also be quickly seen from this screen



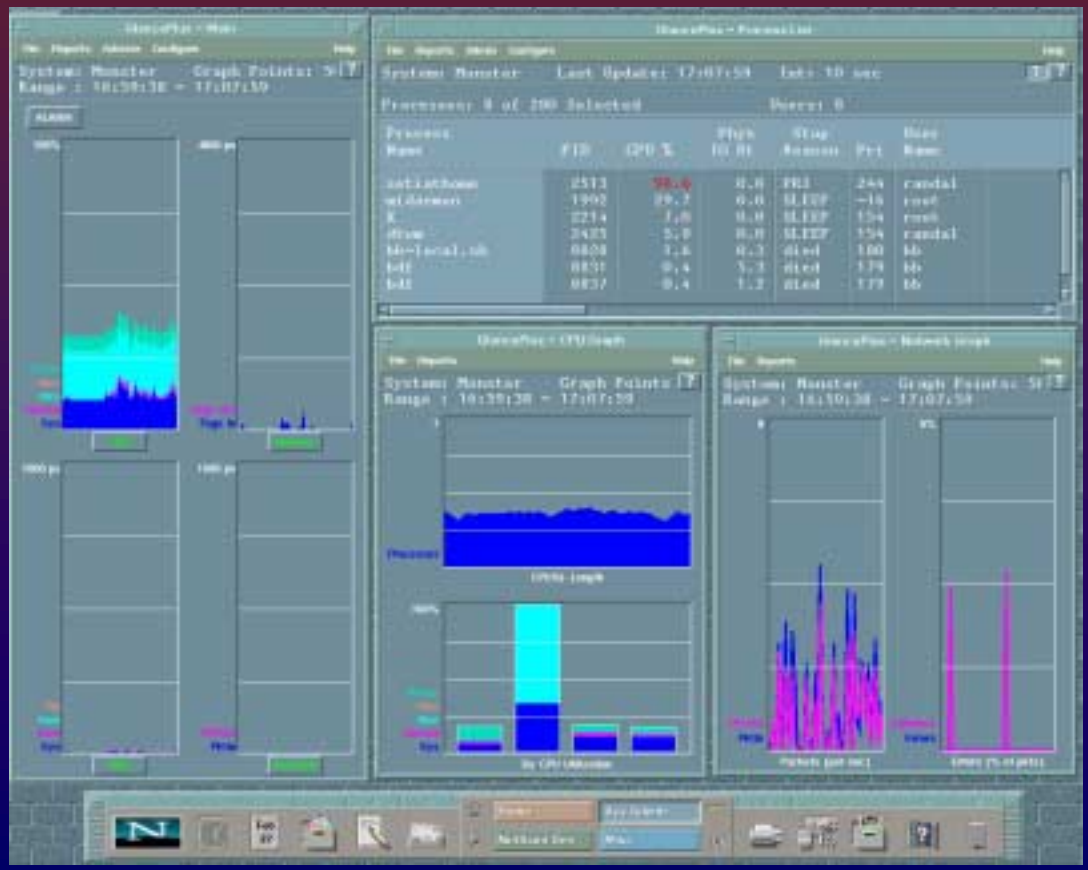
XOSview Screen Shot

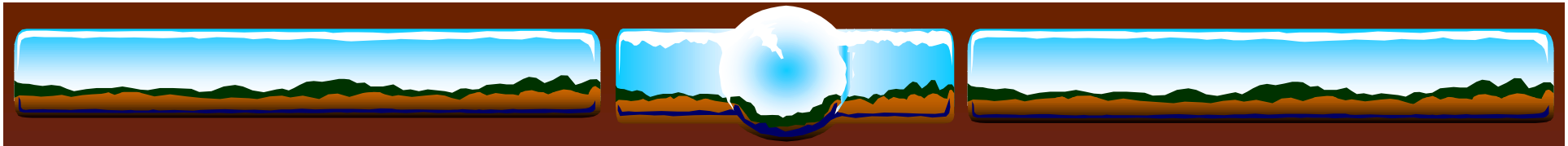
- Xosview (http://lore.ece.utexas.edu/~bgrayson/xosview.html) or RedHat Distribution CDs
- Monitors node health using a X-windows interface



Glance Screen Shot

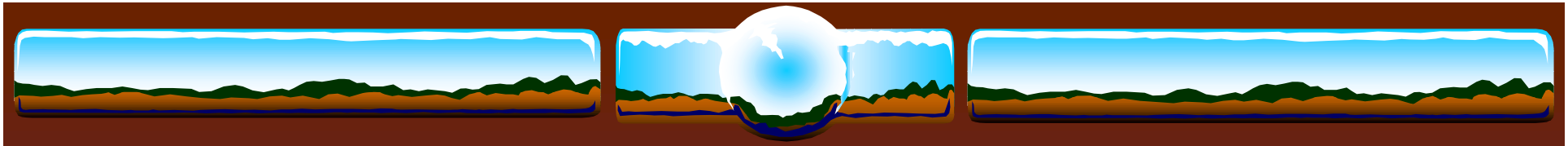
- HP Glance is an add-on product for HP-UX systems. It has a curses and Motif GUI





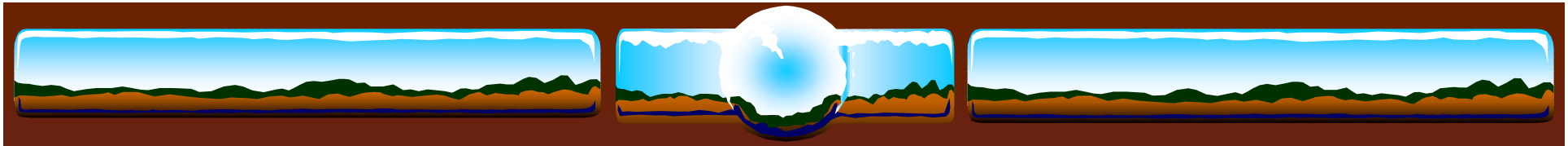
Resource Monitoring

- Load Sharing Facility (LSF) (http://www.platform.com)
- Linux Resource Manager (http://www.arc.unm.edu/~lrm/)
- Maui Scheduler (http://mauischeduler.sourceforge.net/)
- Portable Batch System (PBS) (http://www.openpbs.org)



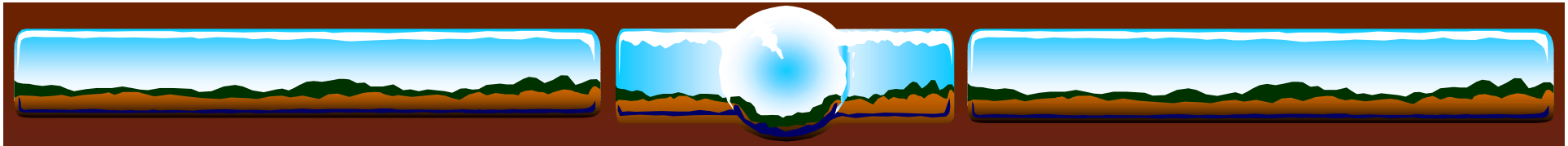
What is LSF?

- A heterogeneous network of computers is viewed as a single system
- Users are no longer limited to the resources on their local workstations
- Programs do not need to be changed or rewritten for use with LSF
- Only a few simple commands are required to submit jobs and execute tasks
- LSF automatically selects the best host to execute a job or task



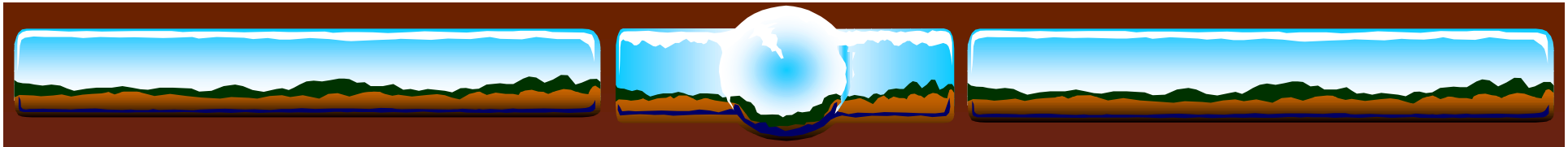
The Complete LSF Suite of Products

- LSF Standard Edition
  - The Standard in Application Resource Management
- LSF Parallel
  - Application Resource Management for Parallel Computing
- LSF Make
  - Make processing for Heterogeneous Distributed Computing
- LSF JobScheduler
  - Job Scheduling for Heterogeneous Distributed Computing



The Complete LSF Suite of Products

- LSF MultiCluster
  - Application Resource Management for Enterprise Computing
- LSF Analyzer
  - Graphical Performance Analysis Tool for Application Resource Management
- LSF Client
  - Extending Application Resource Management to the Desktop

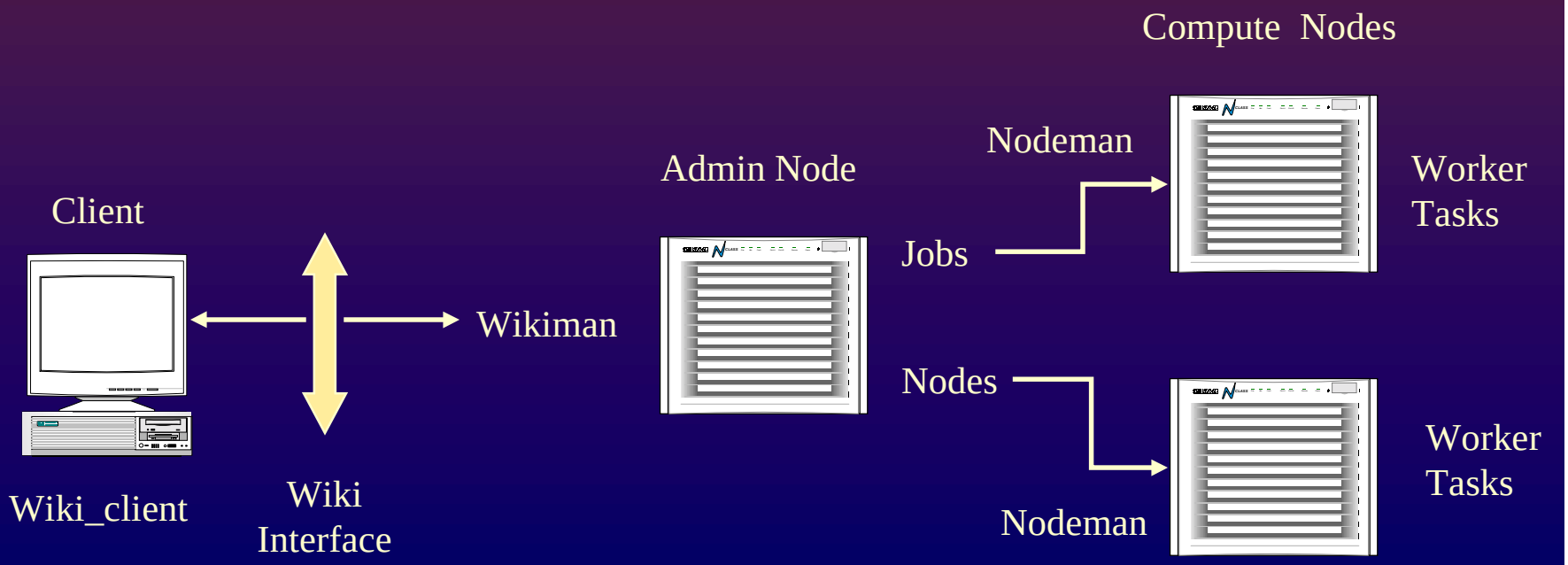


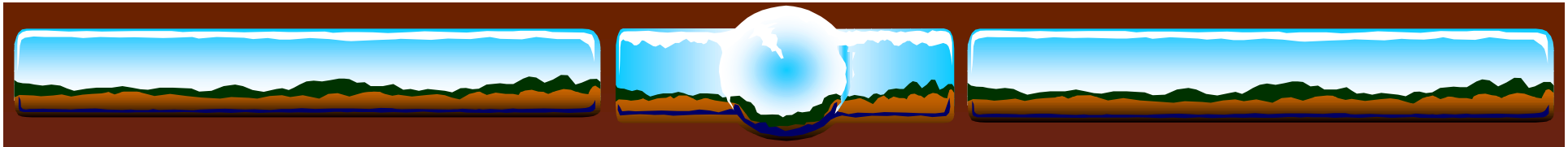
Linux Resource Manager

- History Lesson
  - The Maui scheduler was originally developed to be dependent on the IBM SP Load-Leveler API
  - As interest in the Maui Scheduler for Linux, IRIX, HP-UX, and Windows NT grew it was necessary either to write the interfaces to existing Resource Managers, or to develop a Resource Manager (RM) specifically for the Maui Scheduler
  - In 1998, a Maui High Performance Computing Center team started to develop a generic resource manager called Wiki, from what was the Wiki RM
  - The Linux Resource Manager development began at the Albuquerque High Performance Computing Center and is known as the Linux Resource Manager

Linux Resource Manager

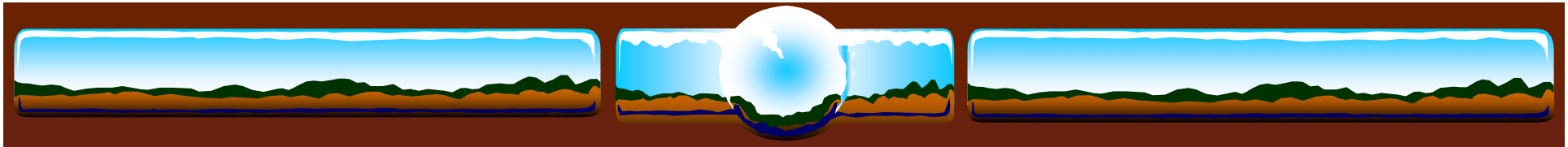
- Nodeman process runs on each compute node
- Wikiman, Maui Scheduler process runs on admin node
- Wikiman gathers state of compute nodes from Nodeman
- Wikiman communicates with Maui Scheduler
- Wikiman launches MPI job on the compute nodes





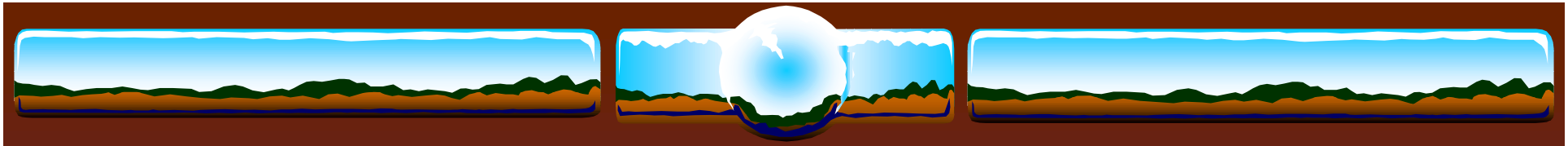
Linux Resource Manager

- *showq, shownode, showjob* are used to view queues, processes running on a node, and to view the resources consumed by a particular job
- Use *submitjob* `program` to submit a job
- Use *canceljob* `program` to cancel a job



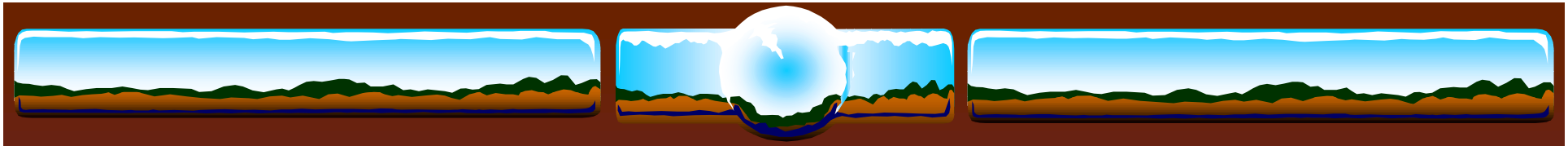
PBS

- Developed for NASA by Veridian Systems
  - Veridian maintains a professional version on a wide variety of platforms
  - More information at http://www.pbspro.com/
- xPBS graphical interface and command line interface
- Job priorities
- Job Interdependencies
- Cross-system scheduling
- Security and ACLs
- Job Accounting
- Automatic load-balancing
- Fully configurable
- Supports MPI, PVM, HPF
- Automatic file staging



Maui Scheduler

- Maui Scheduler is an advanced reservation based High Performance Computing batch scheduler
- Supported on SP, O2K, and Linux/UNIX clusters
- Can be used to extend the functionality and improve the efficiency of sites utilizing the PBS and Loadleveler batch systems



ServiceControl Manager

- Centralized
- Multi-system
- System deployment and recovery
- Multi-system software management
- System configuration management
- Hardware and software fault management
- For more details see the presentation by Humberto Sanchez titled “An Overview of the SCM and the New 2.0 Features”