Managing Linux Clusters

Overcoming the Complexities Associated with Managing Linux Clusters

HP WORLD 2003
Solutions and Technology Conference & Expo

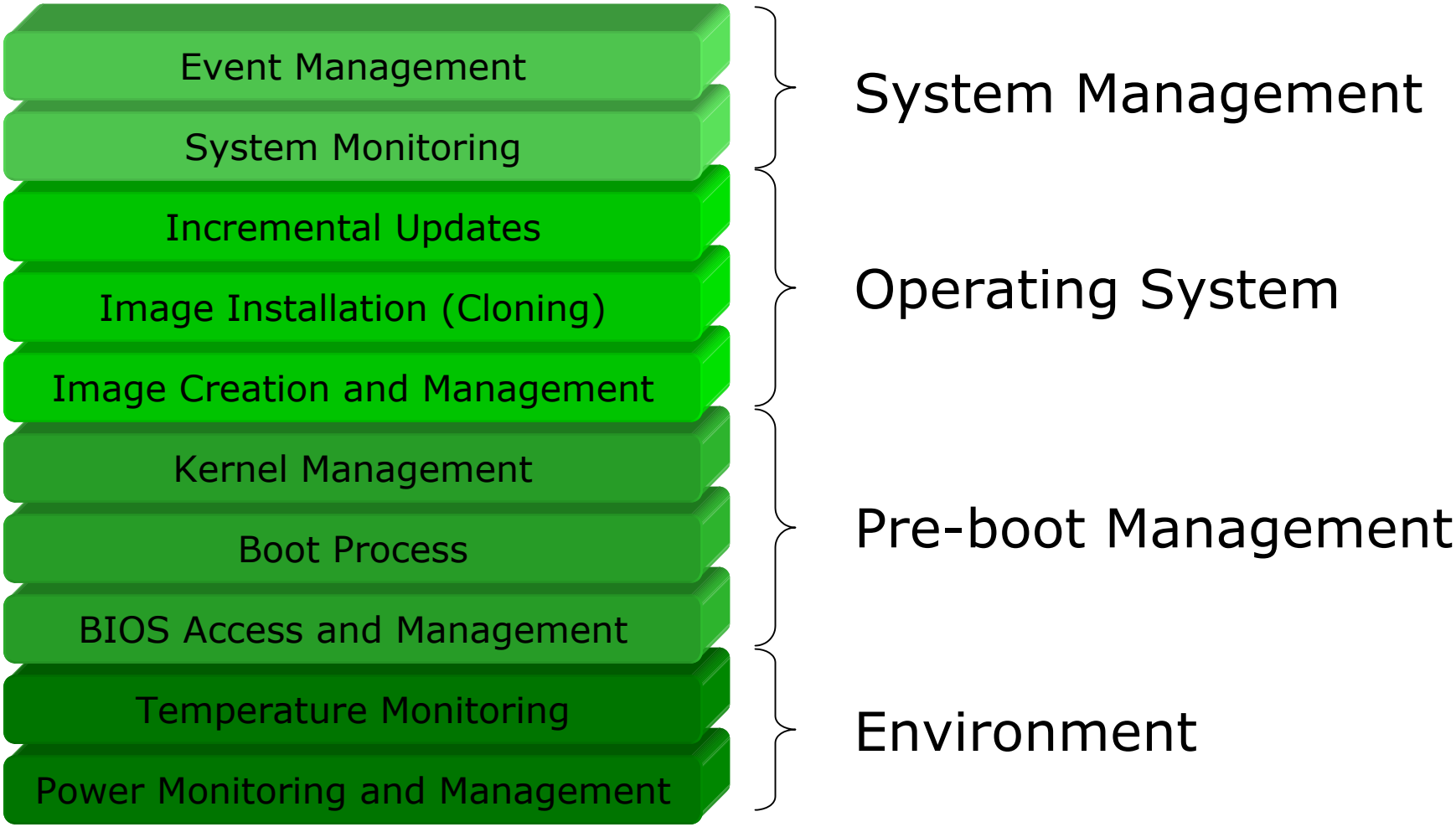
Clusters: What and Why?

- What?
  - Collection of computers networked together to perform a particular application in parallel
- Why?
  - Scalability
  - Cost-effective

Key Management Issues

- Nodes do not contain:
  - Floppy disk
  - CD-ROM
  - Monitor, Keyboard, or Mouse
- Each node has a full/partial installation (even system images)
- Everything must be done over network or serial
- Installation and configuration of the operating system and applications
- Controlling every aspect of the system from a single administration point
- Being able to administer the system remotely

Management Layers



Power and Temperature

- Distributing power evenly
- Controlling power to individual nodes remotely
- Managed via Network or Serial
- Allows “broadcasted” power commands
- Must be scalable
- Monitor internal CPU and chassis temperatures
- Temperatures must be remotely viewable
- Temperature warning system to prevent overheating

What does the BIOS do?

The BIOS is the firmware in charge of:

- Initialising hardware (Memory, Hard Disk, PCI Devices, I/O Ports, etc...)
- Checking for bad memory
- Initialising the boot processes (including the installation mechanism for the OS).

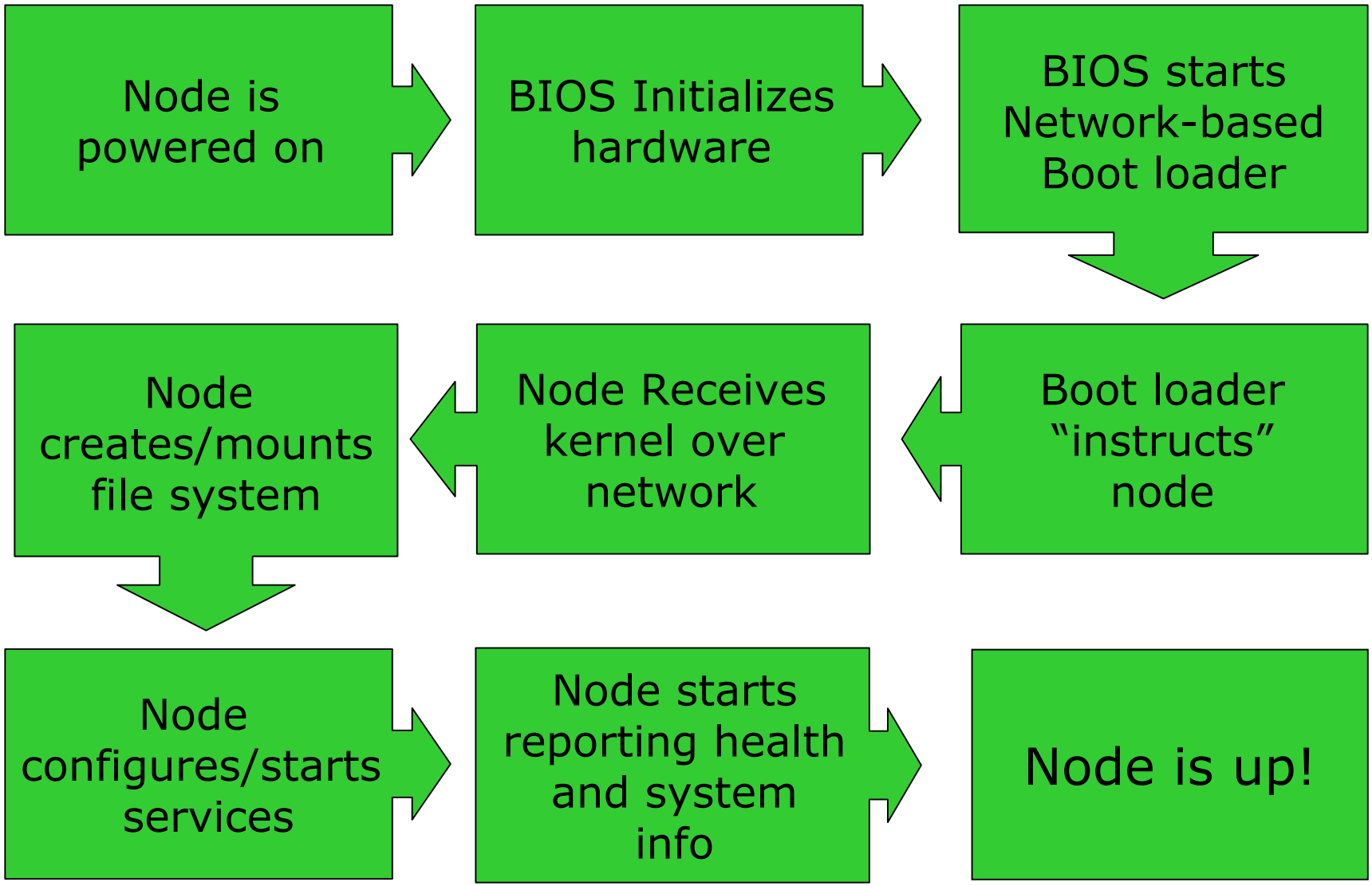
LinuxBIOS Features

- An Open Source BIOS Alternative
- Boots very quickly (3-5 seconds)
- Is remotely accessible
- Configured specifically for cluster systems

Why use LinuxBIOS?

- Speed – LinuxBIOS can increase the boot time of a node by 10-20x
- Accessibility – LinuxBIOS is accessible from within Linux user space – making it accessible while the machine is running from network or serial
- Maintainability – Using LinuxBIOS tools, an entire cluster system can be upgraded in minutes
- Flexibility – Make changes to the BIOS or tools to support custom options (Example: Booting over Myrinet)

Boot Process Overview



Kernel Management

- Nodes should always get their kernels over the network.
  - Simplifies kernel management
  - Easy to upgrade the kernel on the entire cluster
  - Support for optimized kernels
  - Support for packaged kernels or built-from-scratch kernels
  - Support for multiple kernels
  - Kernels can be assigned to node-types
  - Kernels are not tied to OS installation (image)

What is an Image?

- Operating System
- Applications
- Custom Libraries
- Source Code
- Node configuration files (/etc directory)
- Running services
- File system information (partition table or mount points)

Image Creation and Management

- Support for multiple images
  - Images are stored on the “host” machine
  - Images are compressed as large files
  - Uncompressed images can be edited simply by making changes to images root (/) directory
    - Support for chroot
    - Ability to install new RPMs into the image
- Base Linux Install
  - Support for multiple OS versions
  - Support for RPM packages
  - Support for Source files or copied directories

Boot Process Overview

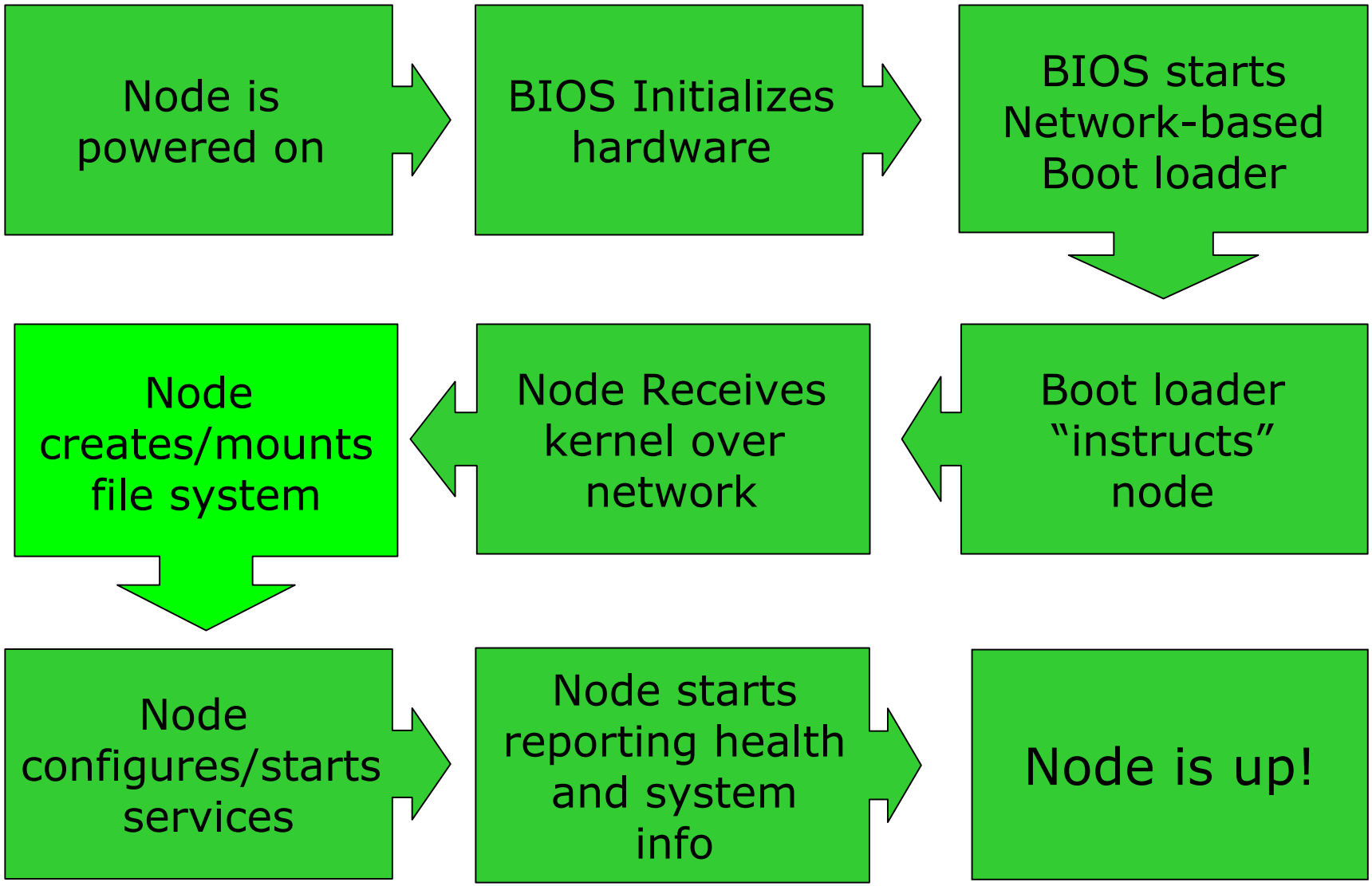
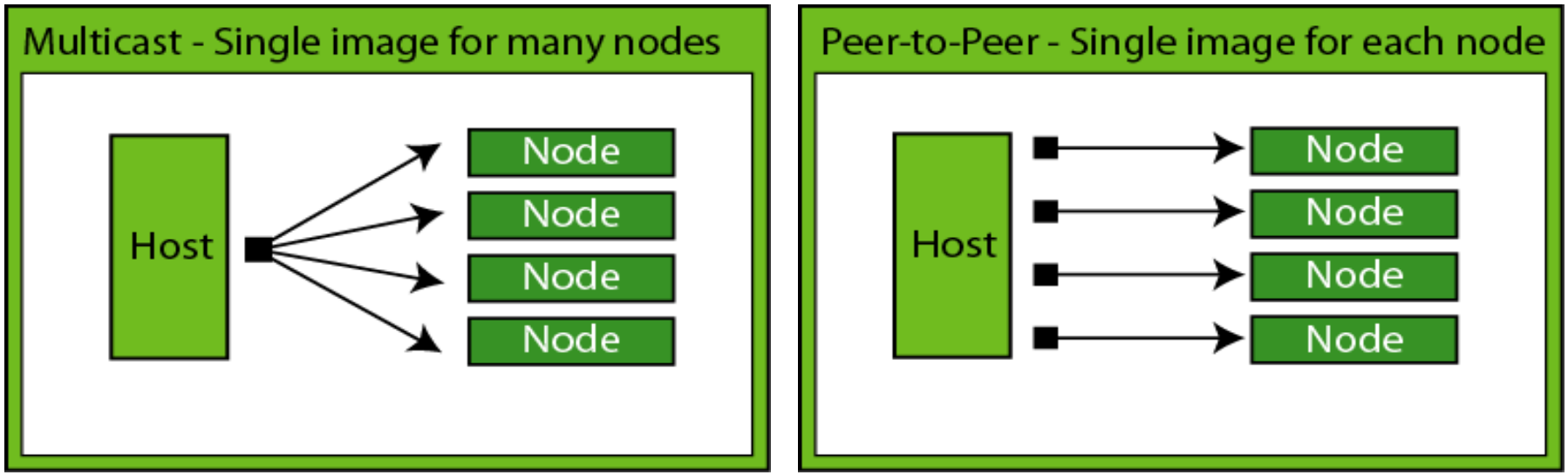


Image Installation

- Cloning (or Provisioning) is the process of installing the Image onto each of the nodes
- Support for different image types:
  - Clone boot image
    - Repartitions the disk
    - Rebuilds the file systems
    - Installs the Image
    - Configures services
  - HD boot image
    - Mounts the previously installed file systems
  - NFS root image
    - Mounts a remote NFS file system as its root file system
    - Allows local scratch and swap

Multicast Technology



- Multicast will clone with 99.9% reliability, that's 999/1000 nodes.
- Multicast will clone an entire system all at once, no matter the size.
- Previously, cloning was limited to about 40-60 nodes at a time.
- Multicast technology is limited by having switches that support multicast.
- Nodes can also be booted and given their kernels using multicast

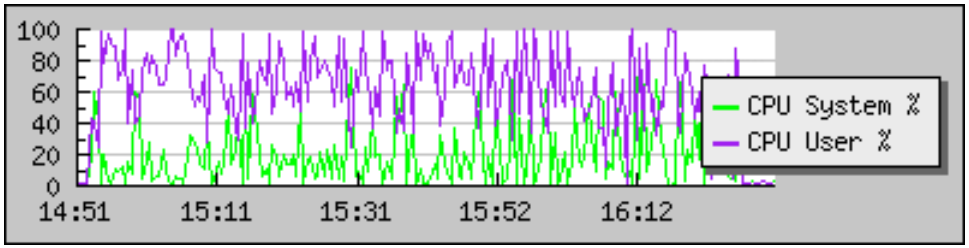
Incremental Updates

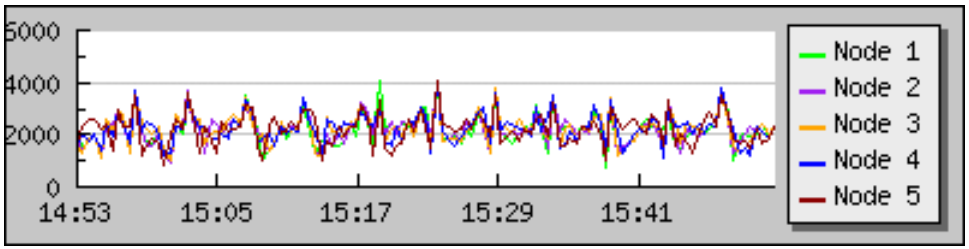
- Parallel Shell & Parallel Update commands
  - Parallel update can take advantage of multicast for large files
- Parallel shell is used for mass administration all/groups of nodes
- Changes to the node should be done in the clone image, other wise the information will be lost the next time the node is cloned.
- Important data should never be stored on the node
- OS should be “disposable”
- Changes to NFS Root nodes are made just by making changes to the NFS root image

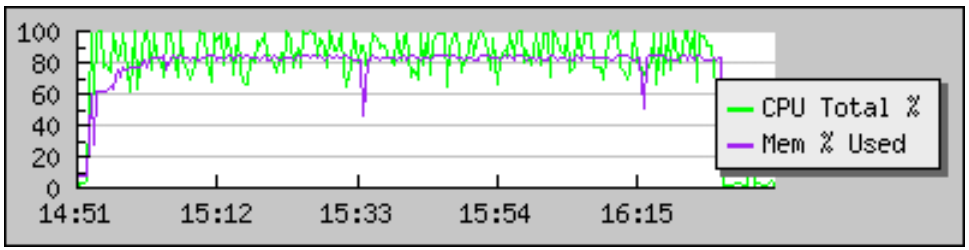
System Monitoring

- Monitoring of each individual system
  - CPU
  - Memory
  - Network
  - Disk
  - Minimal Application
  - OS
  - Node Information (read at start-up)
- Provide monitoring information to Event Engine
- Store monitoring information for historical analysis

Historical Tracking





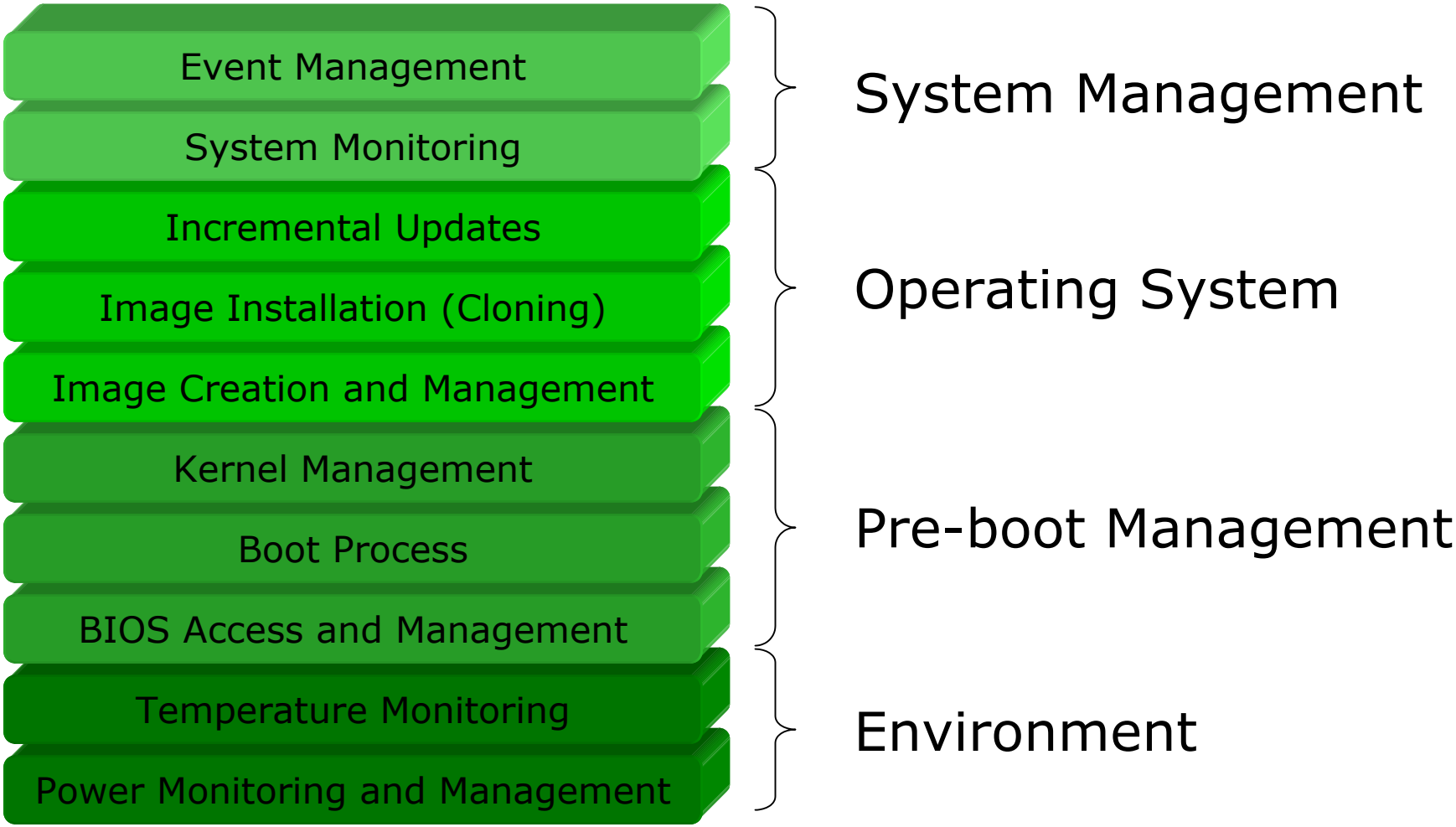


- Displays monitoring information over a period of time.
  - View Cluster performance over time
  - Compare nodes or monitoring values
  - Predict future computing needs
  - Spot system bottlenecks
  - Improve cluster efficiency

Event Management

- Events perform automatic system administration tasks
- Example:
  - If (Temperature >= 60 °) { Shutdown the node; }
  - If (LoadAvg >= 4) { Run a script; }
  - If (RootPartition >= 95%) { Run a script; }
- Send email or take action when an event has been triggered.
- Monitor any default or custom properties
- Support for custom scripts (bash, perl, sym-link)
- Sometimes simple events management can be handled at the hardware level

Management Layers



Software Solutions

- Generally management software is limited to a single tier (System Management, OS & Kernel, BIOS & Environment)

System Management

- Graphing Tools
  - RDDTool, Ganglia
- High-end SNMP based (not great for clusters)
  - Open View, Tivoli, Unicenter, BMC
- Cluster Specific
  - Clusterworx, Beowatch, ABC

Software Solutions

OS Installation

- Redhat Kickstart Based:
  - OSCAR (OSS)
  - xCAT
  - NPACI Rocks (OSS)
- Image Based
  - Clusterworx
  - PowerCockpit

Software Solutions

Power and Temperature (hardware dependent)

- Power
  - Tripp Lite, APC, Baytech
- Temperature
  - LM Sensors, IMPI

Importance of Serial Access

Features:

- Out-of-Band management
- Watch the node boot
- Access/Modify the BIOS
- Universal availability of the serial port on hardware
- Less expensive and more convenient and scalable than KVM.
- Doesn't require X installed
- Support for "serial over IP"
- Supports offline data buffering

Issues to over come

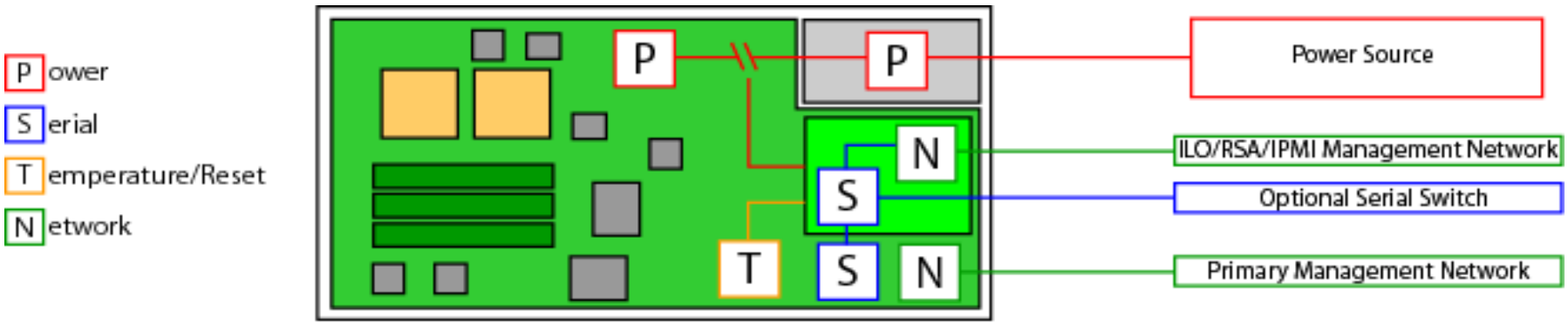
- Poor Scalability
- Low port density (used a lot of rack space)
- Low speed (was 9600, now 115200)
- No remote access
- Focused on Data Center

Serial access solutions

Three types of solution

- Internal Management Cards
  - HP's Integrated Lights Out
  - IBM's RSA
  - Intel's IPMI
- Serial Console Servers
  - Cyclades
  - Lightwave
- Cluster Management Appliance
  - Linux Networx ICE Box

HP Integrated Lights Out



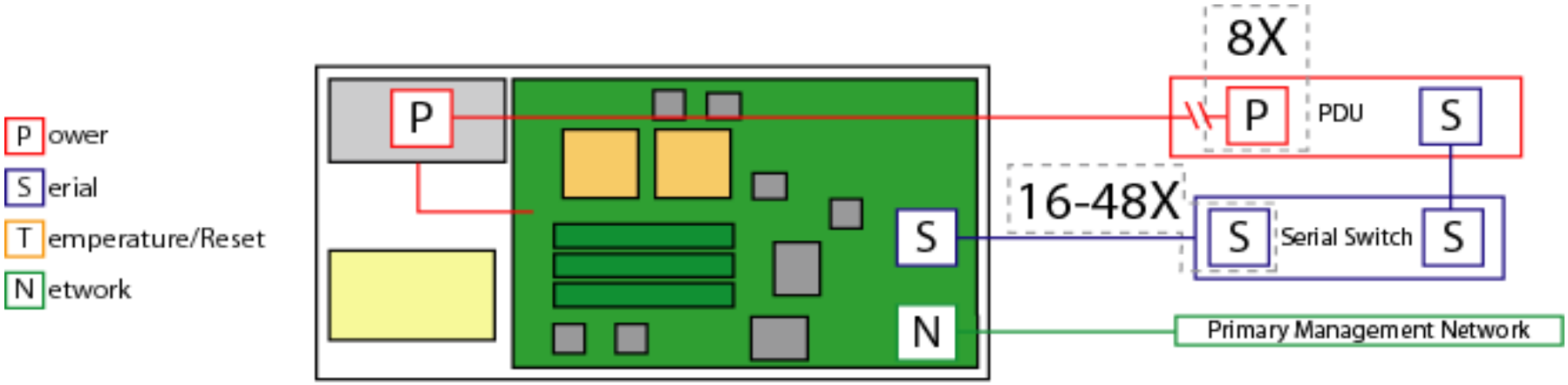
Features:

- Scalability
- Network or Serial Access (or "serial over IP")
- Internal Power management
- Motherboard monitoring
- Onboard event management
- Runs a Mini web server

Drawbacks:

- Scalability
- Second 10/100 network required
- Still requires serial switch
- Very architecture dependent.
- Shell is very limited
- Can't set network settings without DHCP or serial

Serial Terminal Servers



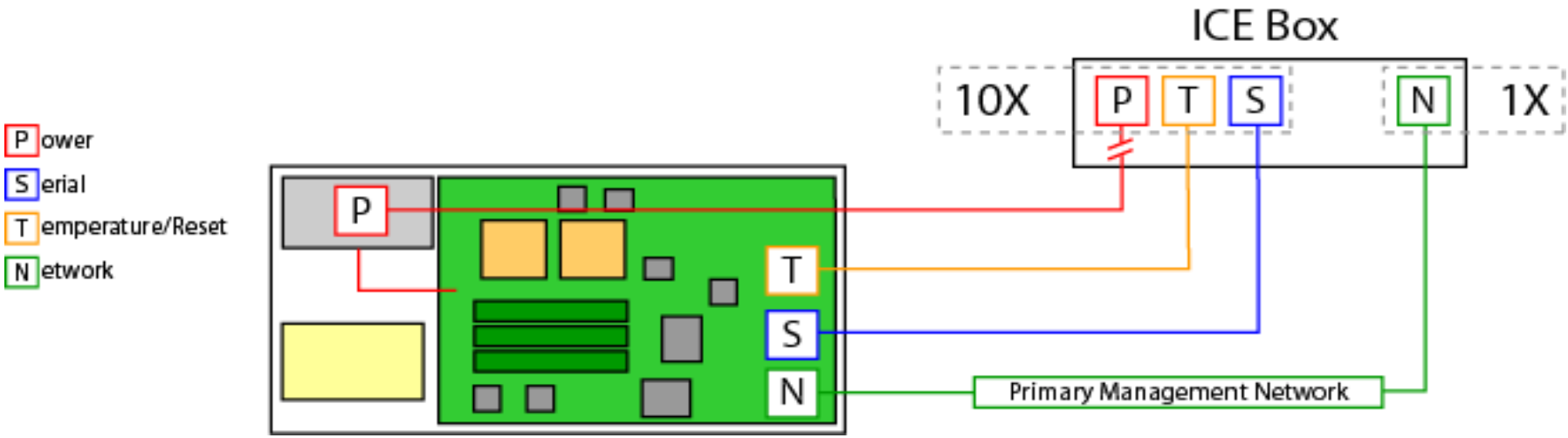
Features:

- Serial Access & Power management
- Up to 48 ports in a 1U space
- Better security features
- Two PCMCIA card slots for expanded functionality
- Web-based management

Drawbacks:

- 15 / 20 Amp power units available with support for 8 nodes or devices
- No temperature, reset, beacon, or other monitoring.
- The PDU requires a serial connection
- Focused on Data Centers

Cluster Management Appliance



Features:

- Serial Access & Power management
- Requires no rack space
- Support for 10 nodes & 2 devices per box – Total of 30 Amps
- Very scalable
- Linux based command shell
- Easy to use Keypad / LCD
- Cluster Specific

Drawbacks:

- No expansion for external devices
- Legacy 10 Base-T Ethernet
- No voltage monitoring

Conclusion

- Questions?