Cray's New Clustering Offering

"A Production Quality Cluster"

John M. Levesque

CUG SUMMIT 2002

Outline

- What's wrong with Clusters?
- What can be done to make clusters more productive?
- Cray's new cluster offering

What’s Wrong With Clusters

- Are they productive
  - What do we mean by PRODUCTIVE
    - Reliable
    - Effective utilization of resources
    - Reproducible
    - Good I/O as well as Compute
    - Hierarchical Storage
    - Good Programming Tools
    - Good Administration Tools

What’s Wrong With Clusters

- Are they productive
  - Most successful clusters are single application clusters
    - Running multiple applications complicates resource sharing

What’s Wrong With Clusters

- Are they productive
  - – The weakest characteristic of clusters is poor, very poor parallel I/O
  - – No one has addressed the problem of archiving cluster files to a SUN, IBM or SGI storage system.
    - NFS is way too sloooow

What's Wrong With Clusters

- Are they productive
  - What do we mean by PRODUCTIVE
    - Reliable NO
    - Effective utilization of resources NO
    - Reproducible NO
    - Good I/O as well as Compute NO
    - Hierarchical Storage NO
    - Good Programming Tools MAYBE
    - Good Administration Tools MAYBE

What Can Be Done To Make Clusters More Useful

- Good resource allocation software
- Good parallel I/O
- Good interface to archival storage

What Can Be Done To Make Clusters More Useful

- Good resource allocation software
  - Checkpoint/Restart
  - Pre-emptive Scheduling
  - Ability to roll out jobs to global file system
    - File system must be good

What Can Be Done To Make Clusters More Useful

- Good parallel I/O
  - Gfs, pvfs and other open source file systems have significant problems
  - Lustre may become the solution
    - Put forth and funded by LLNL
    - Object Orientated file system

What Can Be Done To Make Clusters More Useful

- Good interface to archival storage
  - How to interface to HPSS, Sun’s SAM FS and SGI’s DMF

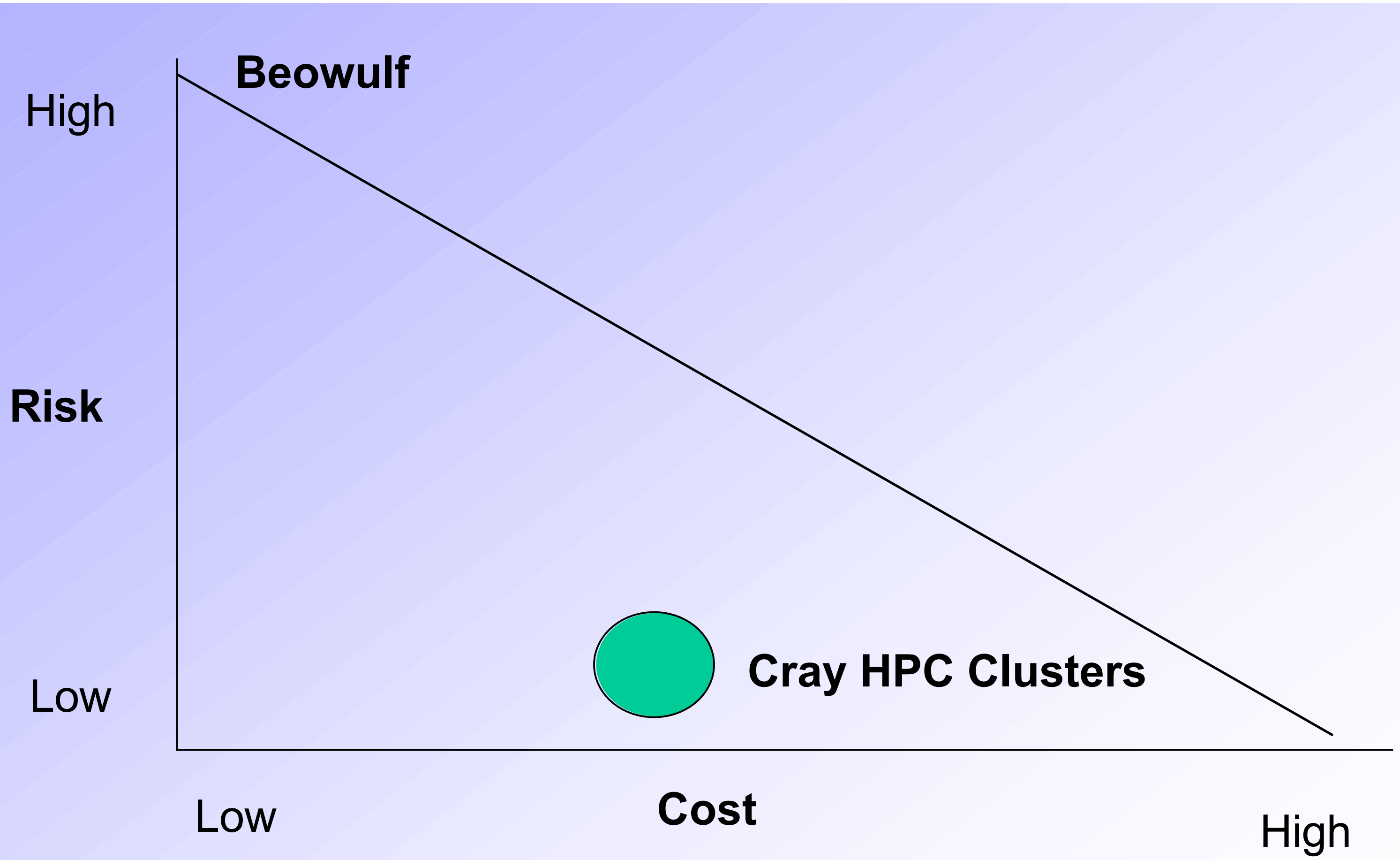
Why Cray

- Of all the HPC vendors, Cray has understood what it takes to make a production quality system
- Cray has much of the software required to supply the functionality to make clusters productive
- Cray has the expertise to port/develop the software for a Linux cluster

Why Cray

- Partnership with the most price conscious server vendor – Dell
- Has the largest concentration in HPC professional expertise
- Has the ability to develop the software to supply production quality HPC clusters

Cray HPC Cluster



Cray’s New Clustering Offering

- COTS based processors Any of several interconnects
- Depending upon customer’s requirements (Myrinet, Quadrics, GigE, etc)
- A robust minimal cluster software stack that will be improved quarterly
- Professional services to design, integrate and maintain the cluster

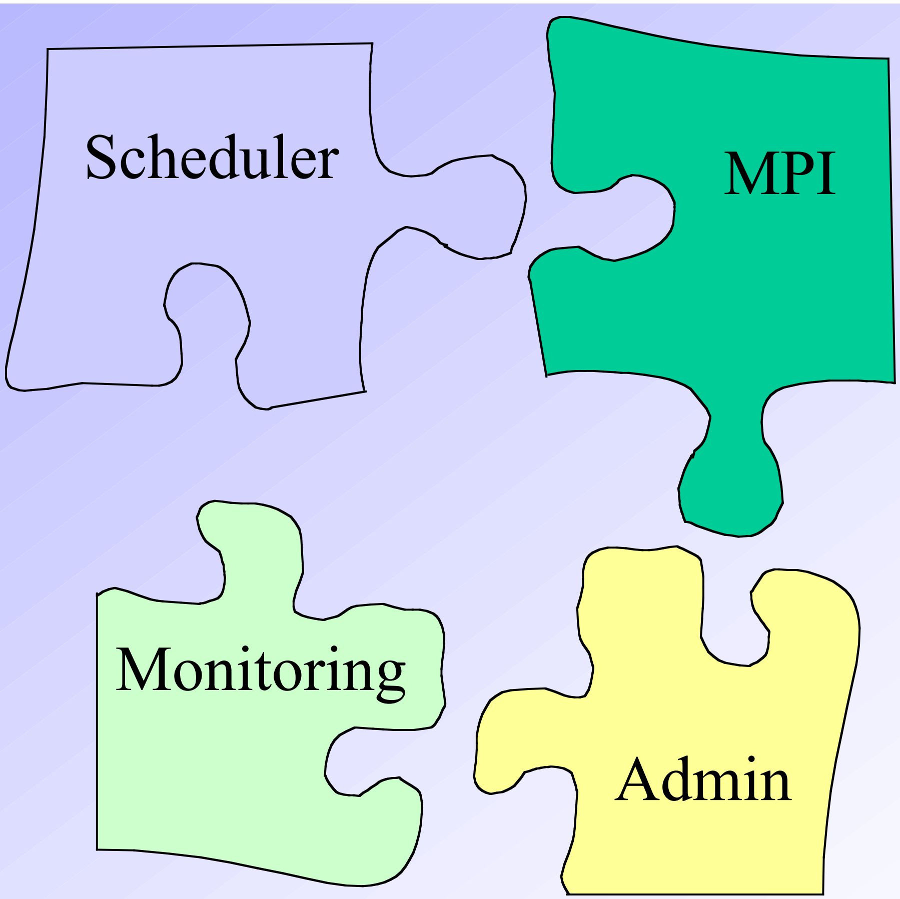
Cray Software Stack

“Robust and Growing”

The Software

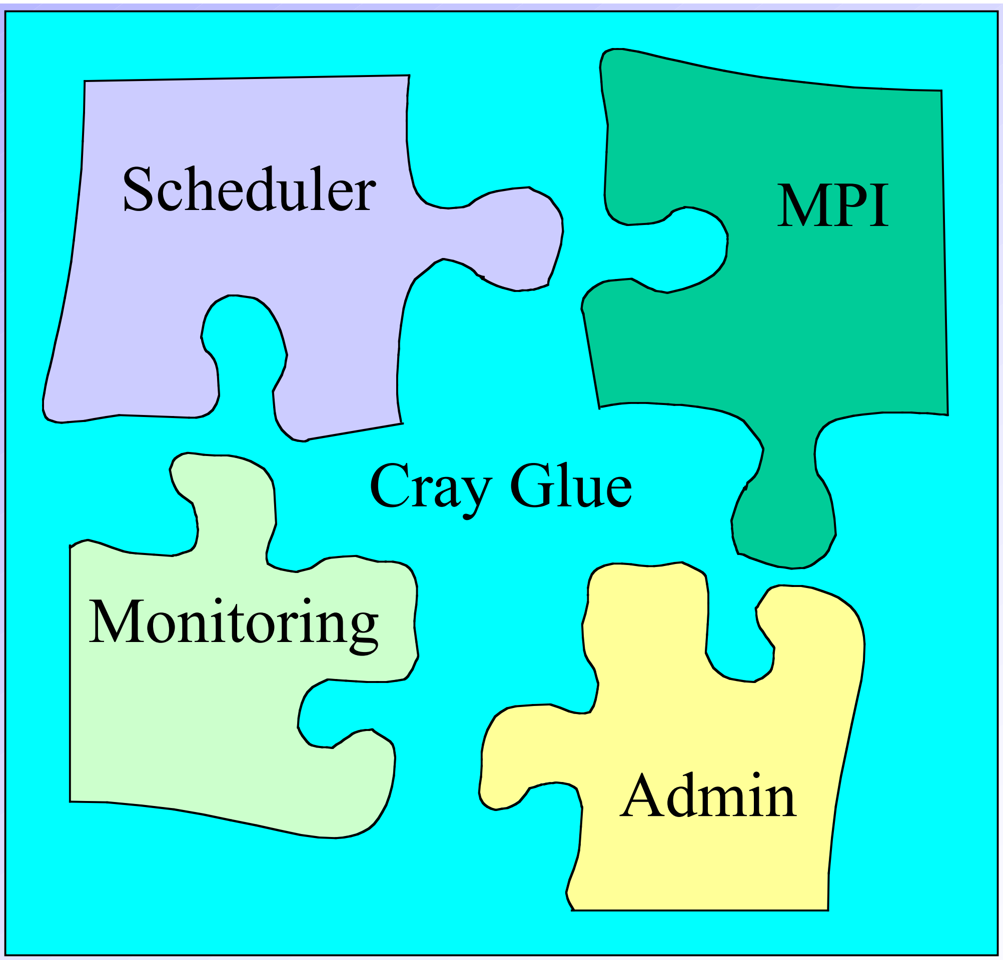
- While the initial offering is no more extensive than other cluster vendors, Cray adds a wrinkle –
  - Cray integrates all the hardware and software – testing to assure that all the components work together and with the customer’s applications (When the customer supplies the apps)
  - As new versions of the software is released, Cray will re-test and re- install a software release that assures that the new releases of the software still works together and with the user’s applications

Available Open Source Software



Everything
fits together.
Kinda…
I guess…

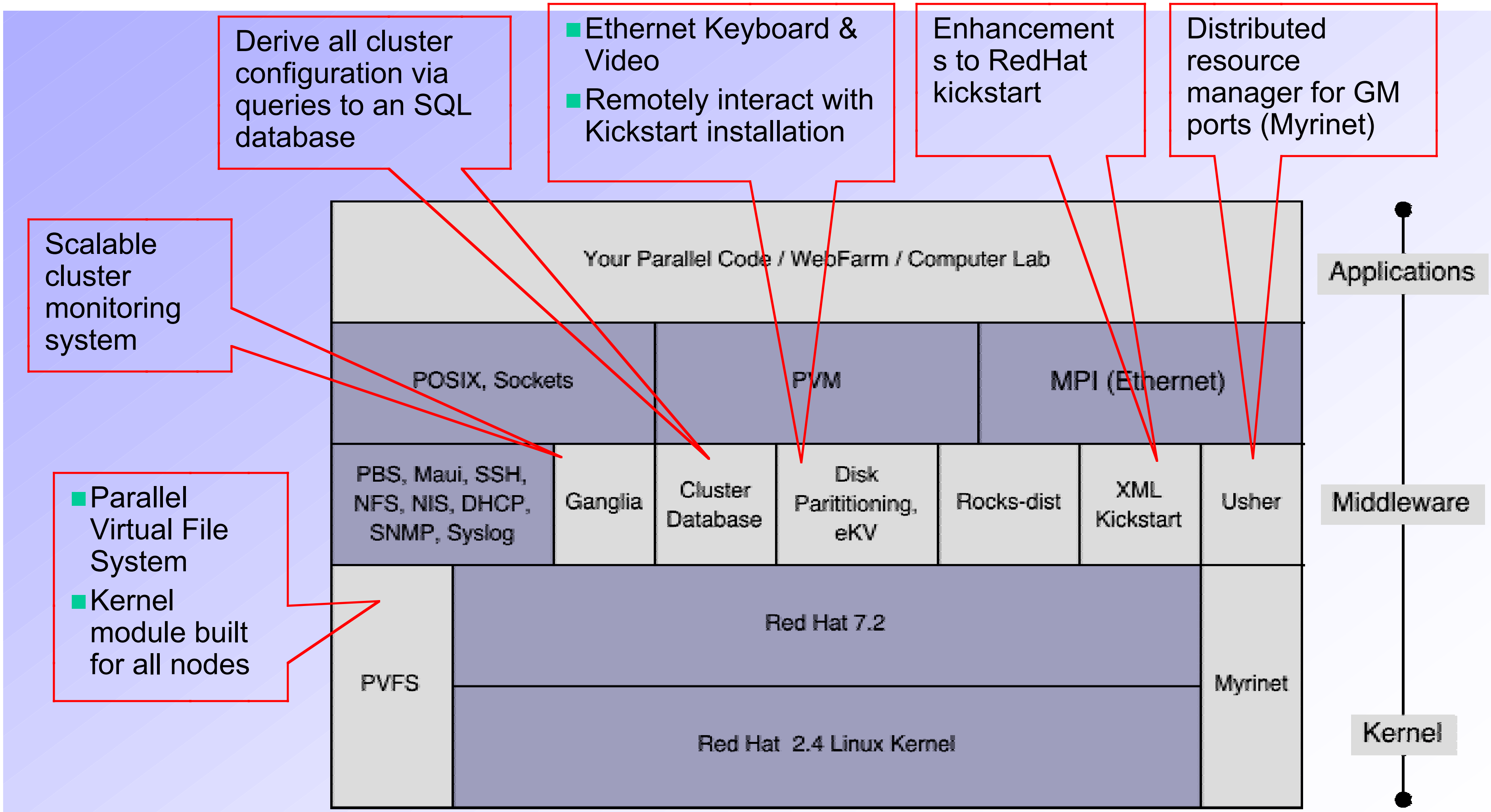
That now works together



Cray's Initial Software Offering

- **Cray's software stack is built upon the NPACI Rocks software. This software is explained in full detail at rocks.npaci.edu. This software contains a significant suite of software components to manage large clusters.**
- Cluster Computing Group at SDSC
  - http://www.sdsc.edu
- UC Berkeley Millennium Project
  - Provide Ganglia Support
- Linux Competency Centre in SCS Enterprise Systems Pte Ltd in Singapore
  - Provide PVFS Support
  - Working on user documentation for Rocks 2.2

ROCKS Major Components



Derive all cluster configuration via queries to an SQL database

- Ethernet Keyboard & Video
- Remotely interact with Kickstart installation

Enhancements to RedHat kickstart

Distributed resource manager for GM ports (Myrinet)

Scalable cluster monitoring system

- Parallel Virtual File System
- Kernel module built for all nodes

Your Parallel Code / WebFarm / Computer Lab

POSIX, Sockets

PVM

MPI (Ethernet)

PBS, Maui, SSH, NFS, NIS, DHCP, SNMP, Syslog

Ganglia

Cluster Database

Disk Parititioning, eKV

Rocks-dist

XML Kickstart

Usher

PVFS

Red Hat 7.2

Red Hat 2.4 Linux Kernel

Myrinet

Applications

Middleware

Kernel

Cray's Initial Software Offering

- **Cray has added numerous scripts and components to interface to Dell's Remote management hardware and software.**
  - Cascaded power up/down. **This allows you to power up/down a set of nodes in sequence, mainly to prevent problems associated with the peak power drawn by a node (or actually a set of nodes) on power up.**
  - Remote Console logging**. This just allows us to keep a trace of what system messages come up on the slave nodes when the cluster in running. This is useful in being able to diagnose problems with individual nodes.**
  - Hardware monitoring utility. **We use this to monitor internal sensors such as fans, thermal sensors, and intrusion latches.**
  - Specific H/W management support for 2650 Dell machines**.**

Cray's Initial Software Stack

- Resource Management
  - PBS with Maui Scheduler
  - LSF also supported
- High Performance I/O
  - PVFS with virtualization software to give redundancy and high performance
  - Lustre
- Performance Tools
  - Intel Compiler
  - Totalview Debugger
  - Vampir MPI Trace facility
  - Application Performance Tools

Professional Services

“The fuel for HPC Clusters”

Will customer pay additional cost?

- Customers are moving to Linux clusters to save money.
  - Many are finding the result is not production quality
- Customers will invest the money they save on hardware on professional services to assure production quality?

Success depends on two complementary offerings

- Value added Software **which provides advanced scheduling, reliability, utility components not available to other Linux integrators**
  - **Make sure that all software plays together on the installed hardware**
  - **Quarterly updates which have been tested prior to installation**

- Professional services **that covers software and extends beyond into applications**
  - **System design will deliver a well balanced hardware system**
  - **Continued on and off site support to assure a continuing production quality system**

Cray's Cluster Offering Tomorrow

- Continue to excel in Custom hardware technology manufactured by Industry leading foundries
  - Supply High bandwidth technology required for the national security and capacity hungry applications
- Enhance COTS technology with hardware and software innovations from custom systems
  - Supply superior HPC systems that compete in the price/performance market
  - Supply Production Quality Clusters
- Continue to grow the HPC professional services
  - Supply superior HPC expertise
  - Supply superior Linux expertise