Cray Operating Systems Road Map

Charlie Carroll, *Cray Inc.*

ABSTRACT:. *This paper discusses the Cray Operating Systems road map. In addition, the rationale for coming changes is discussed.*

KEYWORDS: Operating systems, releases

1. Introduction

The Cray Software Operating Systems and I/O (OSIO) group provides key infrastructure and service components of the software stack.

These components include:

- Compute node kernels
  - XT CNL
- Service node kernel
  - Supports all compute node types
- File systems
  - Lustre
  - DVS (Data Virtualization Service)
- Networking
  - uGNI and DMAPP
  - Portals
  - TCP/IP
- Operating system services
  - Node Health Checker
  - Core specialization
  - Dynamic shared libraries
  - Cluster Compatibility Mode
  - Checkpoint / restart
  - CSA (Comprehensive System Accounting)
- System management
  - Interface to system data
  - ALPS (Application Level Placement Scheduler)
    - Interfaces to PBS Pro, Moab/Torque and LSF
  - Command interface

This paper discusses the main themes to be emphasized in upcoming OSIO software releases, followed by specific features to be delivered in these releases.

2. Release Themes

Upcoming OSIO releases will emphasize certain broad themes. Before getting into specifics, we will take a look at the big picture.

2.1. System stability

Stability and robustness are important to any customer system. They are especially important in large supercomputers with millions of separate components.

Cray has invested substantially over the past two years in defect reduction. By our internal measures, we've made substantial progress in reducing customer bugs. By the most important measure—Software Mean Time To Interrupt (SMTTI)—we've made great progress. SMTTI for large (10+ cabinets) systems has improved from a few hundred hours a couple years go to more than 2,500 hours.

2.2. Performance

Cray's Compute Node Linux (CNL) implementation performs extremely well, largely because we have limited what services and features run on compute nodes.

CLE 3.1 will contain a feature known as core specialization. This optional feature increases the performance of jitter-sensitive applications (generally

those with lots of all-to-all communication). For example, POP has run 30% faster with core specialization.

2.3. Hardware Support

Part of OSIO's mission is to support new Cray hardware as it becomes available. Gemini support will be introduced in CLE 3.1. Gemini, Cray's new interconnect, brings significant performance and reliability features to the marketplace.

Support for AMD's Interlagos processors will be introduced in Ganges.

2.4. File Systems

Cray supports a variety of I/O models. Much of our installed base uses direct-attached Lustre file systems. Through Cray's Custom Engineering group, Cray ships and supports external Lustre file systems. CE integrates white-box servers, Lustre server software and storage. Cray supercomputers connect to these external Lustre servers with Infiniband.

Third-party vendors may be offering Lustre appliances. These self-contained units combine server, software and storage.

Cray supports other file systems through DVS (Data Virtualization Service). DVS projects the file system from Cray service nodes to the Cray compute nodes. The service nodes, in addition to serving DVS, are also clients to the remove, projected file system. This enables applications running on the compute nodes to access, through DVS, the external file system. To date, Cray has used DVS to interface with Panasas and GPFS.

3. Upcoming Releases

Cray will release CLE 3.1 (code-named Danube) in June 2010. There will be three updates to Danube: UP01 in 3Q10; UP02 in 4Q10; and UP03 in July 2011.

Cray will release Ganges in 1H11. This release will support Interlagos, accelerator hardware and SLES11 SP1.

3.1. CLE 3.1 (Danube)

CLE 3.1 will support the Gemini interconnect (for those systems with this hardware). Gemini brings a number of advantages:

- System resiliency in the face of link outages
- Lower latency, 1.7us for nearest neighbour
- Higher injection rates
- Improved error handling

Forest Godfrey of Cray is presenting a paper at this conference in which he explains in details the resiliency features of the Gemini interconnect.

Core specialization allows one core on a multi-core node to be dedicated to overhead (interrupts, system calls, etc.). Jitter-sensitive applications can see a performance benefit from rearranging the work in this fashion. For example, POP got 30% faster when run with core specialization on a large system. Because not all jobs benefit, this feature is specified on a job-by-job basis.

Cluster Compatibility Mode (CCM) allows many ISV applications to run unmodified on Cray systems. CCM works by, at job launch time, creating on the nodes allocated to the CCM applications a cluster-compatible environment. MPI runs over TCP/IP over the high-speed network. Standard services such as ssh, rsh, nscd and ldap are made available. When the job finishes, the services are torn down and the nodes are returned to the regular compute node. In this way, CLE 3.1 systems can run a mix of large-scale, capacity applications intermixed with smaller ISV applications.

Lustre 1.8 will be released in CLE 3.1. Lustre 1.8 includes features to speed up failover (Cory Spitz of Cray is presenting a paper at this conference on imperative recovery); adaptive timeouts; OST pools; and OSS read cache support.

DVS (Data Virtualization Service) will support stripe parallel mode. This allows a file to be spread across multiple DVS servers. In addition, DVS will support server failover in CLE 3.1.

3.2. CLE 3.1 UP01

Much of the focus in UP01 will be, as in past updates, on bug fixes. Shortly after Gemini's introduction, as the hardware and software stack mature, we expect significant improvements in UP01.

UP01 will also contain some features. This is a departure from recent practice prompted by feedback from our customers. Rather than save features for a single annual release, they will go out in periodic updates. This gets features to customers sooner, as well as being easier administratively.

UP01 will have a more scalable RSIP implementation. Each server will be able to support about 5x clients than with base CLE 3.1

UP01 will support Repurposed Compute Nodes (RCN). This allows service node functionality that doesn't require external connectivity—Mom nodes, in particular--to run on compute node hardware. This feature is both more efficient (compute nodes are less expensive than service nodes) and easier to administer (at boot time, the administrator can change the number of repurposed compute nodes). In UP01 RCN will only be supported with Moab/Torque.

In addition to bug fixes, UP01 will also contain performance optimizations, particularly for the Gemini stack.

3.3. CLE 3.1 UP02

UP02 will bring CLE 3.1 to Cray XT4 and XT5 systems. Both are based on the SeaStar interconnect. XT4 and XT5 customers will obviously not get the Gemini-specific features, but they will get the network-independent features.

With UP02, ISV applications run under CCM will natively access the high-speed network. This will boost performance of the ISV application.

DVS support of Panasas will be officially released in UP02.

Checkpoint/restart will be ported from our SeaStar implementation (based on BLCR). This will be available in UP02.

4. Conclusion

This paper has presented specific features which will be coming in 2010 and 2011 releases of Cray's operating system. In addition, we have discussed the themes and thought processes behind our plans.

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency under its Agreement No. HR0011-07-9-0001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

The author would like to thank his colleagues and development team at Cray. Their commitment to producing the world's best supercomputers makes it a pleasure to come to work every day, as well as making this paper possible.

About the Author

Charlie Carroll is Director, OS and I/O with Cray, Inc. If you have comments on our road map, he would love to hear from you at charliec@cray.com.