Cray Operating Systems Road Map

Charlie Carroll, *Cray Inc.*

ABSTRACT:. *This paper discusses Cray's plan to address a number of issues in the system management tools provided with Cray XT systems. Cray will create a single API through which customers can access and affect XT system data.*

KEYWORDS: Mazama, XT, system management, ALPS

1. Current Issues

Managing a supercomputer such as a Cray XT system requires that the system administrator has access to a large amount of diverse system data. Cray's current XT tools for doing so have a number of issues.

- Data are disseminated throughout the system. There is no single, centralized location for accessing and changing system data.
- Customers must use a variety of commands and APIs to access system data.
- XT command usage is restricted to particular nodes in the system.
- The boundaries of Cray's system administration—what is under Cray management and what is not--are unclear.
- The current tools will have scaling issues as system sizes grow from today's sizes.

2. Single Interface

Cray will create a single API through which system administrators and others can access and change system data. On Cray's side of this interface, Cray will maintain a data set which describes the state of the system. This information will include at least application IDs running on the system, job IDs, the hardware configuration of the system, topology information, logical attributes, log files recording system information, system images, partition information and much more.

From the client side of this interface, system administrators will be able to access this information. The existing XT command set will continue to be available, possibly with some rationalizations. Data will also be available in XML format. This industry-standard format allows system administrators to write tools and scripts for manipulating system data in site-specific ways.

There are many advantages to consolidating XT system data behind a single, well-documented API

- Data appear to be in one place. This is much simpler for the administrator. In practice, the data will be in a variety of stores in a variety of formats. This complexity is hidden from users.
- Cray will continue to improve how we handle system data. Because these changes are behind the documented API, customers are insulated from them. Administration tools will work across changes made behind the API.
- XML is an open standard. Delivering system data means that we can extend the data schema such that existing tools continue to function as before. Having an open started creates opportunities for collaboration.

In the coming months, Cray will write a white paper documenting the details of this API. Contact the author for details.

3. Command Ubiquity

On current XT systems different commands are only available from different system locations. For example,

some commands can only be executed from the SMW while others can only be initiated from the boot node. This, of course, is not convenient and hampers administrator productivity.

Cray plans to "wrap" commands with an infrastructure that allows any command to be executed from any location (provided the right networking and security arrangements have been made by the site administrator). This will be more productive for system administrators.

4. Boundaries of Cray's System Management

Some customers have externalized their Lustre servers while others have externalized their login nodes. This has created some confusion about what system management tools should manage and not manage.

To clarify, Cray will:

- Manage directly hardware built by Cray;
- Monitor white boxes serving XT functions such as Lustre and login;
- Make information available via SNMP so that Cray XT systems can be managed by existing data center tools.

5. Scaling Issues

Cray expects to see within a few years systems approaching 1,000,000 cores. This is substantially larger than today's systems and will pose challenges for many areas. The specific challenge within system administration is managing the large quantity of data. Imagine a command which accesses a database once for each core. At small system sizes, this approach is manageable. With a million cores, the shortcomings will quickly be apparent.

Cray plans to take two actions (and undoubtedly others as we learn more) to mitigate scaling issues in system management.

Cray will cache system data in clever ways in order to reduce database accesses. In the example above, we may be able to replace a million database queries with a much smaller number. While the functionality is hidden behind the API, the performance difference will be clear.

Cray will create filters to allow more focused inquiries into the system data. For example, the API will support a call to list all the nodes in a system along with the status of each. This could be a lot of data. If the administrator is interested only in inactive nodes with 8GB of memory, a filtered query which returns only those nodes will greatly reduce the amount of data flowing through the system.

6. Conclusion

This paper has defined five problems present in Cray's current XT system management infrastructure, along with our plans to solve each. These features will appear in future Cray XT releases.

Acknowledgments

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 system management direction, he would love to hear from you at charliec@cray.com.