Cray XC System Level Diagnosability: Commands, Utilities and Diagnostic Tools for the Next Generation of HPC Systems

Jeffrey J. Schutkoske
Cray Administrative Environment (CAE)
Cray, Inc.
St. Paul, MN, USA
jjs@cray.com

Abstract— **The Cray XC system is significantly different from the previous generation Cray XE system. The Cray XC system is built using new technologies including transverse cooling, Intel processor based nodes, PCIe interface from the node to the network ASIC, Aries Network ASIC and Dragonfly topology. The diagnosability of a Cray XC system has also been improved by a new set of commands, utilities and diagnostics. This paper describes how these tools are used to aid in system level diagnosability of the Cray XC system.**

Keywords-Cray XC, diagnostic, diagnosability, computer architecture

I. INTRODUCTION

System Diagnosability is a suite of software tools designed to provide Cray field support and end customers a tool chain to quickly and reliably identify hardware and software problems in the Cray XC system.

System Diagnosability is not just about diagnostics. Diagnostics are just one aspect of the tool chain that includes BIOS, user commands, power and thermal data and event logs. There are also component level tests that are used to checkout the individual components, but quite often issues do not appear until substantial scale (20+ cabinets) is reached. From experience over the last few years, we have seen that no single tool or diagnostic can be used to identify problems, but rather multiple tools and multiple sources of data must be analyzed to provide proper identification and isolation of hardware and software problems.

System Diagnosability of Cray XC system components include the Aries High Speed Network (HSN), Intel compute processors, Intel Xeon PHI co-processors, Nvidia GPUs, Memory, Intel Quick Path Interconnect (QPI), Peripheral Component Interconnect Express (PCIe), Cray cabinet power and cooling, as well as, storage solutions and connectivity to storage solutions.

This paper will focus on three areas of the Cray XC system as follows:

1. Aries High Speed Network (HSN)
2. Compute processors and co-processors
3. Cray XC cabinet power and cooling

There are a number of areas that Cray needs to address in the future as follows:

1. Additional workload tests
2. End-to-end storage tests including Infiniband
3. Diagnostic data analysis tools
4. HSS dashboard

II. SYSTEM DIAGNOSTICS

System Diagnostics are used to validate the various components individually or concurrently in the system. The system diagnostics focus testing in a number of areas as follows:

- Boot
- Confidence
- Stress
- Performance
- Workload
- Error and Data Reporting

Boot level tests perform tests at boot time or under special diagnostic kernel images. A boot level test includes Built-In Self Test (BIST), tests within BIOS or off-line diagnostics that execute under an off-line diagnostic kernel image.

Confidence level tests perform tests that validate the functionality of the component. The confidence tests shall validate different test algorithms, data patterns and sequences. A confidence level test focuses on Aries Block Transfer Engine (BTE) or Fast Memory Access (FMA) tests but not both BTE and FMA concurrently. It also focuses on the processor, memory, co-processor or the Hardware Supervisory System (HSS).

Stress level tests perform tests that are designed to maximize stress on the hardware. These tests are deliberately intense or thorough testing used to determine the stability of a given system or entity. It involves testing beyond normal operational capacity, often to a breaking point, in order to observe the results. Reasons can include: - to determine breaking points or safe usage limits; to confirm intended specifications are being met; to determine modes of failure (how exactly a system may fail), and to test stable operation of a part of the Cray XC system outside standard usage.

Performance level tests perform tests that can measure and calculate the performance of a given component. These tests are also configured to have an expected performance level for the given component to compare actual results against. On the Cray XC systems typical components where performance is tracked include the Aries Network point-to-point performance, node memory read/write performance, node QPI performance, node computational performance,

node to GPU or co-processor PCIe bandwidth performance, and co-processor or GPU computational performance.

Workload level tests are tests that either simulate a generic application workload, benchmark or are actual applications used to verify that they system is ready to execute user applications.

Error and data reporting is critical to system diagnosability. Capturing warnings, faults and errors, as well as, thermal, power and status data from all devices into a central location on the SMW allows for better system analysis of problems. Along with logging the data, providing commands to parse and decode the data is essential to proper analysis.

III. ARIES HIGH SPEED NETWORK

The Cray XC system High Speed Network (HSN) connects each processing node to a single Aries network interface (NIC). Since there are four NICs on each Aries ASIC, there are four processor nodes connected to each Aries. Each Aries has 4 PCIe links and 10 optical ports. There are 96 Aries ASICs in a full sized optical group, composed of 16 per chassis times 6 chassis.

The Dragonfly topology is a hierarchical network consisting of two layers of a flattened butterfly topology. The first layer is a two dimensional flattened butterfly that connects all of the Aries ASICs with an optical (local) group. The optical group refers to a pair of cabinets or six chassis.

The first dimension within the optical group is the "green" dimension that connects the 16 Aries ASICs within the chassis. The second dimension within the optical group is the "black" dimension that connects the six chassis within the two cabinet optical group. The optical ports are the "blue" dimension that connects the optical groups within the Cray XC system. The five network ranks correspond to traversing green, black, blue, green and black link in order.

A. Aries Boot Tests

The Hardware Supervisory System (HSS) executes an Aries Memory Built-It Self Test (MBIST) as part of the Aries initialization. The Cray Intel processor BIOS performs the initial PCIe lane training and reports the PCIe lane width, link speed and status to the boot process.

There are a number of off-line Aries diagnostic test suites executed during manufacturing, at time of system installation and during certain scheduled Preventive Maintenance (PM) times. These test suites are defined as follows:

- Aries ASIC Functional Tests (AFT)
- Cabinet Functional Tests (CFT)
- System Functional Tests (SFT)

The AFT suite independently validates each ASIC including its input, output and functionality. The CFT suite is used to test out the Aries router in the Aries ASIC and the interaction with the other Aries routers in a cabinet in an off-line test mode. The SFT suite is used to test out the top level of the Aries ASIC and the interaction with the rest of the system and other Aries ASICs in the system in an off-line test mode using an off-line diagnostic Linux kernel. These tests execute in an off-line environment, meaning that CLE is not executing on the node or nodes.

B. Aries Confidence Tests

The System Stress Test (SST) HSN on-line diagnostics are designed to stress the various components of the Cray XC system. This requires large amounts of data to be generated, written to, read from and compared to expected results. Each targeted functional area within the Cray XC system is tested. Each test performs a key assessment of the functional area using a unique test algorithm. A key assessment for a functional area is defined as a functionality test, data compare test, accessibility test, a component-level metric or performance test. For each test that computes a component-level metric a deviation assessment compares the deviation from the nominal reference value against the deviation tolerance. If the key assessment fails the test reports the failure and a component report is generated for each failure. If a component-level deviation assessment fails a component report is also generated. A component report includes the location of the component in the system, the expected metric value, the actual metric value and the reference value.

The SST HSN on-line diagnostics validate the Aries Block Transfer Engine (BTE) and Fast Memory Access (FMA) transfer types in the SST environment.

This suite of individually selectable tests directly tests each BTE transfer type. The SST BTE tests are as follows:

- SST BTE All to All Test (*xtbte_ata*): The BTE All to All Test ensures that all logical endpoints go to all other end points using BTE Put and/or Get transactions.
- SST BTE All to One Test (*xtbte_ato*): The BTE All to One Test ensures that all logical endpoints target one end point using BTE Put and/or Get transactions. A round robin approach is used to step through each end point in the configuration.

This suite of individually selectable tests directly tests each FMA transfer type. The SST FMA tests are as follows:

- SST FMA All to All Test (*xtfma_ata*): The FMA All to All Test ensures that all logical endpoints go to all other end points using FMA Put and/or Get transactions.
- SST FMA All to One Test (*xtfma_ato*): The FMA All to One Test ensures that all logical endpoints target one end point using FMA Put and/or Get transactions. A round robin approach is used to step through each end point in the configuration.
- SST FMA Atomic Memory Operations (AMO) Test (*xtfma_amo*): The FMA AMO test is an All to All diagnostic test using all AMO types. It tests AMO Put transactions and AMO Get transactions.

C. Aries Stress Tests

This suite also provides an FMA and BTE concurrent test, *xtfbc*. The test is designed to test the dedicated FMA and BTE logic blocks concurrently, while stressing the shared hardware like the Processor Interface (PI), Network

Interface (NICs), Netlink (NL), network tiles and high speed links. The FMA and BTE threads exchange data between like thread types as well as FMA to BTE and BTE to FMA thread data exchanges. All threads have the ability to synchronize globally to keep all threads/ranks operating together as a single system test.

These tests all support partitioning and nearest neighbor mode.

If multiple ranks per node are selected, a thread is started for each rank on the node and each thread is mapped to a core. The diagnostic allocates a buffer for the diagnostic and a buffer for each additional node and rank under test. For performance considerations, the diagnostic statically allocates the required buffer space to support the number of cores and ranks selected. The available memory per node limits the number of Aries Network ASICs that can be tested at any given time.

It is not recommended to attempt to run all cores within the Cray XC system, unless the Cray XC system is very small (less than 4 cabinets). The FMA and BTE SST diagnostics can saturate the bandwidth from the node processor to the Aries Network ASIC at around 3 to 4 cores depending on the core processor speed, memory size and memory bandwidth.

D. Aries Performance Tests

The *xta2a* test is used to measure the performance on all-to-all communication for sets of nodes corresponding to the physical structure of an XC30 system: blades, chassis, groups and the whole system. The test is designed to run on as many nodes as are available, reporting variation in performance over sets of nodes of a given size. For example, to run 512 instances of a blade level test on 2048 nodes and report variation between them.

The *xta2a* test generates a high network load. In particular it stresses the PCIe interfaces that connect each node to Aries. Poor performance on this test correlates well to high rates of PCIe errors (logged on the SMW). The *pcitest.sh* wrapper script exercises this use of *xta2a*. The test executes on all nodes for a period of ten minutes. This is sufficient to detect nodes with rates of PCIe errors that impact application performance. The *pcitest.sh* script reports start and end times for the test. These times are used with the SMW command, *xtpe*, to select errors reported in the same interval.

The *xta2a* test uses MPI_Alltoall with the DMAPP optimizations enabled. As such it is representative of a real application. The *xta2a* test does not require dedicated access to the whole system. It can be run on a subset of nodes allocated by the batch system. The impact on performance of other applications is low if all nodes in an electrical group are allocated to a test. The *pcitest.sh* script can be adapted to create a batch script suitable for this purpose.

E. Aries Error Reporting

Advanced Error Reporting (AER) is enabled in the Cray Compute Node Linux (CNL) kernel by default. With AER enabled the Aries NIC logs PCIe errors to the root complex on the Node side. This is also true for both the Nvidia Kepler/Atlas GPUs and Intel Xeon PHI Knights Corner co-processor PCIe errors. The CNL kernel reads these PCIe errors and logs them in the node console log, which is saved on System Management Workstation (SMW). It also reports them via the Hardware Error Log Channel connected from the node to the Blade Controller (BC), which forwards them to the SMW Hardware Error Logger Daemon (*xthwerrlogd*). The PCIe errors are viewable using the SMW command, *xtpcimon*.

The Aries System Stress Tests (SST) HSN on-line diagnostic tests can be executed periodically as batch jobs or interactively from the login nodes to validate system functionality between customer application job runs. The SST on-line diagnostic tests execute in the Cray Linux Environment (CLE). The SST on-line diagnostic tests execute in user space using the uGNI library, the Generic Hardware Abstraction Layer driver (GHAL) and the Generic Network Interface driver (GNI) to access the Aries Network ASIC. These diagnostics do not use Message Passing Interface (MPI) or Distributed Memory Application (DMAPP) API.

IV. Compute Processors and Co-Processors

The Cray XC system provides 3 different types of compute configurations including dual socket Intel processor (Intel SandyBridge or Intel IvyBridge), single socket Intel processor with an Nvidia GPU (Kepler or Atlas) and single socket Intel processor with an Intel Xeon PHI co-processor Knights Corner (KNC).

A. Processor BIOS Tests

The Cray provided BIOS initializes the Intel processor. The Cray BIOS also initializes, trains and reports on various hardware components as follows:

- QPI bus
- Memory DIMM
- Aries PCIe bus
- Co-Processor PCIe bus
- PCIe device bus

The Cray BIOS reports any DIMM failures during memory training including any MCA errors that were detected. Memory size and speed are verified and reported. It also reports all QPI and PCIe link width, speed and status. Any errors are logged and reported to the SMW command, *xtbounce*, as part of the system boot sequence. The node is prevented from booting if any hardware failure is detected during BIOS execution. The Cray BIOS logs are copied to the Cray SMW on any BIOS error or failure for further analysis.

B. Intel Processor Diagnosability

There are a number of diagnostic tests available for the compute nodes. The CPU stress test, *xtcpuburn*, provides a computationally intensive CPU test to heat up the processor. The test can be run for a specified period of time and a specified number of passes. If, during this time, a calculation error is detected, the core routine aborts and an error is

reported. Initially each processor core within the node runs a copy of the test. The main function starts a unique copy in a thread assigned to the processor. The testing is synchronized.

The memory test, *xtmemtester*, targets all of the processor memory. *xtmemtester* is an effective user space memory test for stress-testing the memory subsystem. It is very effective at finding intermittent and non-deterministic faults. The maximum amount of memory that *xtmemtester* can test is less than the total amount of memory installed in the system; the operating system, libraries, and other system limits take some of the available memory.

The NUMA test, *xtnumatest*, is a set of tests that exercise and test the NUMA capability of a Symmetric Multi-Processors (SMP) node. *xtnumatest* is a group of tests that verifies that each processor core within an SMP is able to allocate and access memory that is local to the socket and across the QPI to memory on the remote socket. The remote memory in the context of this test does not include memory connected to other SMPs that may be accessible via the Aries network. *xtnumatest* also provides a suite of tests that stress the node by exercising all cores simultaneously to stress the memory paths from each CPU to local and remote memory. Finally it also provides a performance test to validate the QPI performance.

The Intel processor stress test, *xtcpudgemm*, provides a computationally intensive processor tests to validate the Intel SandyBridge or IvyBridge processor. This test outputs the performance and the power for the processor during each pass of the diagnostic test. This test uses the standard CBLAS DGEMM. It also validates the results of the DGEMM matrix multiply.

The Intel In-Target Probe (ITP) is a JTAG bus with some Intel-specific signals and protocol added. It is traditionally used to debug software and diagnose Intel processor (SandyBridge or IvyBridge) problems. Typically, this is done with an Intel ITP interface connected to the processor via the eXtended Debug Port (XDP). Cray has implemented an embedded ITP so that no external hardware needs to be connected to the Cray XC system. The embedded ITP is used as a processor hardware debug tool. Python bindings exist for the ITP library and several python scripts have been written to take advantage of this feature. These scripts reside on the SMW and are available via the SMW command *xtitp*.

Many of the scripts provide useful hardware debug information about the PCIe configuration and status, QPI configuration and status, processor information, MCA errors, MSR data, and the package Power Limit (turbo) registers. Executing this command on the SMW temporarily pauses the processor, until the data is read from the processor and resumes the processor once the read is complete.

The CNL kernel reads MCA errors (correctable and uncorrectable) and logs them in the node console log, which is saved on System Management Workstation (SMW). It also reports them via the Hardware Error Log Channel connected from the node to the Blade Controller (BC), which forwards them to the SMW Hardware Error Logger Daemon (*xthwerrlogd*). The SMW command, *xthwerrlog*, displays all hardware errors logged via the Hardware Error Channel including MCA errors. The SMW command, *xtmcadecode*, decodes and provides detailed explanation for Intel MCA errors.

C. *Nvidia GPU Processor Diagnosability*

There are a number of diagnostic tests available for the compute node with Nvidia GPUs. The Cray GPU DGEMM test, *xkdgemm*, is a standard double precision floating point matrix multiply application. It utilizes CUDA to execute the matrix multiply on the Nvidia GPU and uses the standard CUDA BLAS library. The Cray GPU Memory test, *xkmemtest*, is used to test the GPU memory for hardware errors and soft errors using CUDA. The Cray GPU PCIe bandwidth test, *xkbandwidth*, is used to measure the memory copy bandwidth of the GPU. It can measure device-to-device copy bandwidth, host to device copy bandwidth for pageable and pinned memory, and device to host copy bandwidth for pageable and pinned memory.

The Cray GPU stress test, *xkstress*, performs stress and performance test across nodes. This test includes STREAMS, GEMM, and PCIe tests. It is an MPI application that compares performance results across the blades, cabinets and system.

The Cray GPU check application, *xkcheck*, validates the Nvidia GPU hardware configuration, firmware version, CUDA Driver API, NVML API and CUDA runtime comparisons. The application can display the Node IDs and cnames for each delta found. It is an MPI application that compares the results across the blades, cabinets and system.

There is also a single node version of HPL, which is used as the workload test for the Nvidia GPU. The majority of the processing is all done on the GPU. There is no sharing of data between Nodes or GPUs.

D. *Intel Co-Processor Diagnosability*

There are a number of standalone Intel Xeon PHI co-processor diagnostic tests. The Cray PHI DGEMM test, *xtphidgemm*, is a standard double precision floating point matrix multiply application. The test also outputs the KNC PHI temperature, memory temperature, and power usages on each pass. The Cray PHI PCIe bandwidth test, *xtphibandwidth*, measures the PCIe bandwidth and memory bandwidth of the Intel MIC KNC. Measurements include device-to-device copy bandwidth, host to device copy bandwidth, and device to host copy bandwidth. Random floating-point data of a default size of 128 MB is verified after each copy.

There are a number of standard workload tests that are also available for the Intel PHI co-processor. The GEMM benchmark is based on the Intel Math Kernel Library's Basic Linear Algebra Subroutines (MKL BLAS) SGEMM and DGEMM operations. These perform the multiplication of two matrices in single (SGEMM) and double (DGEMM) precision. The workload SHOC tests applications measure the PCIe bus bandwidth from host-to-device (BusSpeedDownload) and from device-to-host (BusSpeedReadback) by transferring messages that range in size from 1 KB to 64 MB. The workload STREAM

applications measure sustainable bandwidth for data transfers between off die memory and on die processor cache.

The KNC kernel reads the KNC MCA Detected Unrecoverable Errors (DUE) and Corrected Errors (CE) and logs them in the KNC console log, which is saved on System Management Workstation (SMW). Compute blades with Intel Xeon processors and Intel KNC coprocessors have 2 console streams, one for the Xeon/host and one for the KNC coprocessor. The Xeon host processor processes the MCA error from the KNC coprocessor and sends it to HSS via the same Hardware Error Log Channel that is used between the Xeon host processor and HSS. The SMW command, *xthwerrlog*, displays all hardware errors logged via the Hardware Error Channel including MCA errors. The SMW command, *xtmcadecode*, decodes and provides detailed explanation for Intel MCA errors.

E. PCIe Error Reporting

Advanced Error Reporting (AER) is enabled in the Cray Compute Node Linux (CNL) kernel by default. With AER enabled the Nvidia Kepler/Atlas GPUs and Intel Xeon PHI Knights Corner co-processor PCIe errors. The CNL kernel reads the processor errors and logs them in the node console log, which is saved on System Management Workstation (SMW). It also reports them via the Hardware Error Log Channel connected from the node to the Blade Controller (BC), which forwards them to the SMW Hardware Error Logger Daemon (*xthwerrlogd*). The SMW command, *xthwerrlog*, displays all hardware errors logged via the Hardware Error Channel including PCIe errors.

V. CRAY XC CABINET POWER AND COOLING

The Cray XC liquid cooled cabinet provides a significant number of temperature, velocity, humidity, and water temperature and pressure and power sensors to maintain optimal levels of power and thermal control within the Cray XC system. Cray HSS software polls these interfaces on a consistent frequency to monitor the data in the cabinet.

In addition to the Cray sensors, Intel processors provide a single wire signal for their Platform Environment Control Interface (PECI). Intel uses the Intelligent Platform Management Interface (IPMI) Specification version 2.0, 2004 as the protocol to connect to the Intel Management Engine (ME). This interface is used to monitor thermal, power and electrical conditions of the Intel processor.

Nvidia also provides the capability for Cray to monitor the Nvidia SXM GPU accelerators via an i2c bus using the Nvidia SMBus Post-Box Interface (SMBPBI). This interface is used to monitor thermal, power and electrical conditions of the Nvidia SXM GPU accelerator.

The HSS software installed on the SMW, blade controllers and cabinet controllers monitors the data from the various sensors in the Cray XC cabinet. The Cray HSS System Environment Data Collections (SEDC) reads and logs the thermal, power and environmental data on the SMW. Alerts and warnings are sent to the SMW and are displayed via the SMW command, *xtconsumer*.

To validate the HSS hardware and software, the SMW HSS diagnostic utility, *xtcheckhss*, is used. The SMW command validates the Cray XC system HSS infrastructure by checking each blade and each cabinet. It can be used to get a quick validation that the HSS infrastructure is functioning normally and can also be used to trouble-shoot a blade or a given cabinet. On a blade it validates the basic blade functionality. It can also validate HSS voltages, Aries voltages and current, temperatures, PDC sensors, as well as, the Intel processor and DIMM temperatures and voltages. For a cabinet, it validates the basic cabinet functionality. It can also validate the HSS voltages and rectifiers in the cabinet. It is also used to validate the HSS infrastructure for a compute cabinet, blower cabinet, pre-conditioner cabinet and all air sensors in the compute or pre-conditioner cabinet. Additionally it verifies all of the HSS firmware versions are properly flashed and verifies the HSS Inventory Serial EEPROM (SEEP).

The *xtcheckhss* utility also provides a set of diagnostic tools that test various components within the HSS infrastructure including the rectifiers, blowers and the water valve. These tests change the state of the Device Under Test (DUT) and should not be run during production time.

It also provides a throughput test to the HSS cabinet and blade microcontrollers. This test performs read/write of a pattern of sliding 1's on 0's and then sliding 0's on 1's. It compares the read value with the write value and flags any differences. This test uses a scratch register on the controller and is safe to run during production time.

VI. FUTURE CONSIDERATIONS

Additional work is planned to continue to enhance system diagnosability. One area is to include additional workload tests that have been found to test and stress the Cray XC system in ways that diagnostics have not been able to. While workload tests have been shown to find difficult problems, they generally are not good diagnostic tools. Often times they can only report that "something" has happened that was not expected.

Diagnostics need to continue to evolve and be enhanced to not just test individual components, but to also test the system in similar ways to the workload tests. Future work within diagnostics will focus on understanding how the workload tests execute and stress the Cray XC system and to then enhance the Cray XC system level diagnostics.

Diagnostic tests can also be added to test additional components of the Cray XC system beyond the current Aries Network, compute processors and co-processors and the HSS components. The diagnostic tests need to include end-to-end storage tests including the Infiniband connection to the various storage solutions.

As with any large data problem, Cray can provide enhanced diagnostic data analysis tools that can sift through the data collected on the SMW and make the connections between the data and failures or potential failures on the Cray XC system. Analysis tools can quickly review data over a period of time and present the data of interest to the system administrator.

Finally the data can be presented in an HSS dashboard where operations staff can quickly see where failures are being detected.

VII. SUMMARY

Cray has provided a number of diagnostics, commands and utilities that enhance the system diagnosability of the Cray XC system. The focus of system diagnosability has been on ensuring that each component is functioning properly by ensuring that each component can be validated and that all data is captured. Each aspect of the tool chain has been enhanced on the Cray XC system. Enhancements have included BIOS, SMW commands, utilities, and diagnostics, as well as, power and thermal data and event logs. Future enhancements are planned to continue to improve system diagnosability of the Cray XC system.