Evolution of Cray Management Services

Tara Fly
Cray Management Services
Cray, Inc.
Introduction

Acknowledgements:
- CMS Team
- Service Team: Jeff Caplow, Jeff Becklehimer

Concepts

New Futures

Future Direction
Overview

- Extreme Scalability Mode
- Shared Libraries
- ISV Applications

- Improved Network Resiliency
- External Services
- Hardware Options

- Core Specification / Huge Pages
- Application Reconnect / Relaunch
- MPMD
- Workload Manager

- Improving Availability
- Increasing Use Cases
- Growing Feature Set
Cray System Administration

Administration
• Collecting and isolating failures
• Root Cause Analysis
• Data collection and analysis

System Improvements
• Uptimes
• Feature Set Complexity
• Hardware
Cray Management Solution Response

- Centralized Data Collection
- Standardized Data Formats
- Flexible Dump Collection
- Event Correlation
- Improved System Accounting
- Health Checking
Architectural Premises

- Scalable extensible frameworks
- Capabilities provided through plugins
- Multiple delivery formats
- Iterative development model
  - Base capabilities provided in initial release
  - Add incremental functionality over time
- Customers can extend for specialized use cases
Lightweight Log Manager Architecture
Lightweight Log Manager

- Provides standard logging mechanism for Cray software
- Provides centralized log aggregation on SMW
- Standardizes logging format
- Allows forwarding to an external log host
- Supports configurable log rotation
- HSS controller log forwarding
- Leverages rsyslog

/var/opt/cray/log/p1-20110210t213201/messages-20110210.gz

- base dir
- session id
- log name
- chunk
- comp. suffix (opt.)
LLM Resiliency
Modular System Dump

Python-based xtdumpsys
- Scalable mechanism for fault collection
- Supports parallelism
- Core infrastructure with plugins
- Site extensible
- Scenario support providing targeted data collection
- Significant performance improvements
Modular System Dump Architecture
Performance and Fault Detection

Data collection organized by run level style hierarchy

All plugins at a given run-level can be run in parallel up to a maximum parallel count

Dump scenarios provide targeted collection for specific failure modes
  - HSN congestion,
  - admindown node(s)
  - down node(s)
  - panicked node(s)
  - power and cooling failures

Log window support to gather logs from a specified time range or delta from current time
Node Health Checker Improvements

Reservation-level testing
- Allow site administrators to move testing to the end of a batch-job reservation
- Helps eliminate potential false positives

<table>
<thead>
<tr>
<th>Application Tests</th>
<th>Reservation Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application</td>
<td>Reservation</td>
</tr>
<tr>
<td>Filesystem</td>
<td>Filesystem</td>
</tr>
<tr>
<td>ALPS</td>
<td></td>
</tr>
<tr>
<td>Accelerator</td>
<td></td>
</tr>
<tr>
<td>ugni</td>
<td></td>
</tr>
</tbody>
</table>

Improved CNCU Resiliency
Current Node Health Tests

- Memory test
- Filesystem Test
- Application Exited Check
- Apinit Ping Test
- Ugni Test
- Reservation Test
- Accelerator Test
- Plugin Test
- Hugepage fragmentation test
Node Health Checker

Configuration File Changes

[Options]
advanced_features: on
nhcon: on
dumpdon: off
suspectenable: y

[Memory]
Action: Log
WarnTime: 20
Timeout: 30
RestartDelay: 30
Threshold: 600
Sets: Reservation

Automation as part of installation to convert existing configuration files to new format
Resource Utilization Reporting

- Scalable framework for data collection
- Site extensible plugins
- Multi-stage data collection
  - Data aggregated on compute nodes
  - Data pulled back to MOM nodes
  - Post-processing of data
- Support for multiple backing stores
- Plugins for use cases delivered incrementally
  - GPU Utilization
  - Power Management
  - Application Completion Reporting
  - POSIX Accounting
  - CSA

Andrew Barry “Resource Utilization Reporting on Cray Systems” Thursday May 10th 11:00am
Simple Event Correlation

What is SEC?

Event Monitoring System
Correlates events in area of log analysis, system monitoring, network management and system security

- Deliver SEC RPM with base SMW distribution media
- In collaboration with Cray Service, provide a reference set of rules and installation for site use of SEC
- Reference implementation provides email notification
- Updates delivered by Field Notice between release
- Latest Field package incorporated in subsequent release
Questions ?