Project Caribou; Streaming metrics for Sonexion
Craig Flaskerud
Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

Cray Inc. may make changes to specifications and product descriptions at any time, without notice.

All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, and URIKA. The following are trademarks of Cray Inc.: APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, REVEAL, THREADSTORM. The following system family marks, and associated model number marks, are trademarks of Cray Inc.: CS, CX, XC,XE, XK, XMT, and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis. Other trademarks used in this document are the property of their respective owners.
Cray storage architect since 2013
Oracle Storage 2010-2013 – Storage & Big Data
Sun Microsystems 2001-2010 – Storage Systems
LSC 2000-2001 – SAM-QFS
Let’s talk about Project Caribou!
Why Caribou?

Enable new storage administrative techniques using visual analysis of metric and event data, correlated with workload manager information

For Who?

Storage administrators and Performance Analyst's to easily understand the performance of the Sonexion and it's infrastructure. Operations Staff and be alerted when events and metric thresholds are exceeded.

How?

• Data model
• Data Collection
• Data Integration
• Data persistence management
• UI and Workflows
• Customization and site integration
Caribou Data Model
Dimensions of Caribou Data Model

Time series names begin with cray_storage.* and cray_job.*
cray_job.read_bytes_sec
cray_storage.write_bytes_rate

Time 2017-05-08T16:44:20Z
Region RegionOne
Tennat_id 529612f8fbc5b332b66aab062afdf41
Component lustre
Device OST000a
device_type ost
Hostname unknown
Job_id 4058797
Product snx
Service storage
system_name snx11242
value 29461.066666666666
value_meta
Database Query Example

```
# influx --database mon --host localhost --precision rfc3339 --execute \
"select * FROM /cray_job.write_bytes_sec/ \nWHERE system_name='snx11242' \nAND job_id='4058797' " \

name: cray_job.write_bytes_sec
```

<table>
<thead>
<tr>
<th>Time</th>
<th>Region</th>
<th>Tenant ID</th>
<th>Component</th>
<th>Device</th>
<th>Device Type</th>
<th>Hostname</th>
<th>Job ID</th>
<th>Product</th>
<th>Service</th>
<th>Service Value</th>
<th>Value Meta</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017-05-08T16:44:20Z</td>
<td>RegionOne</td>
<td>529612f8fbc8b4332b6a062afdf41</td>
<td>lustre</td>
<td>OST000a</td>
<td>ost</td>
<td>unknown</td>
<td>29461.066666666666</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017-05-08T16:44:21Z</td>
<td>RegionOne</td>
<td>529612f8fbc8b4332b6a062afdf41</td>
<td>lustre</td>
<td>OST0008</td>
<td>ost</td>
<td>unknown</td>
<td>48251.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017-05-08T16:44:24Z</td>
<td>RegionOne</td>
<td>529612f8fbc8b4332b6a062afdf41</td>
<td>lustre</td>
<td>OST0000</td>
<td>ost</td>
<td>unknown</td>
<td>98330.666666666667</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017-05-08T16:44:26Z</td>
<td>RegionOne</td>
<td>529612f8fbc8b4332b6a062afdf41</td>
<td>lustre</td>
<td>OST0003</td>
<td>ost</td>
<td>unknown</td>
<td>68594.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Collection
The diagram illustrates data flow and integration. Here’s a breakdown:

- **SNMP provider** feeds into the **Syslogd**.
- **SNX Syslog-ng** also connects to **Syslogd**.
- **Perfquery, ibnetdiscover** (provider) connects to **Infiniband provider** and **JobID provider**.
- **Jobevent provider** connects to **Seastream provider**.
- **Lustre & Jobstats** are connected to **Seastream provider**.
- The **Kafka message bus** acts as the central hub for **InfluxDB Consumer** and **Threshold Consumer**.
- **Elastic Search** integrates with **InfluxDB** and **Influx DB**.
- **UI (User Interface)** connects to **InfluxDB** for visualization.

The flow shows a comprehensive data pipeline from various providers to consumers, ultimately leading to alert generation and UI visualization.
- Hardware SNMP Events
- Syslog
- IB counters & Topology
- Lustre stats
- Sonexion node stats
- Jobstats
Data Integration and Transport
● Kafka
● Pub/Sub topics and partitions
● JSON messages
● Syslog
Data Persistence
● Kafka consumers
● Logs in Elasticsearch
● Time series in InfluxDB
● Threshold engine
User Interface and Workflows
## Jobs

<table>
<thead>
<tr>
<th>ID</th>
<th>User</th>
<th>Job Name</th>
<th>Start Time</th>
<th>End Time</th>
<th>File Size</th>
<th>IOPs</th>
<th>Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>377678</td>
<td>1356</td>
<td>growfiles_mpi</td>
<td>2017-05-09 05:16:49</td>
<td>2017-05-09 05:21:16</td>
<td>963.0kB</td>
<td>1.2kB/op</td>
<td>497.1k</td>
</tr>
<tr>
<td>377700</td>
<td>1356</td>
<td>growfiles_mpi</td>
<td>2017-05-09 06:01:23</td>
<td>2017-05-09 06:05:41</td>
<td>970.5kB</td>
<td>1.6kB/op</td>
<td>501.6k</td>
</tr>
<tr>
<td>377677</td>
<td>1356</td>
<td>growfiles_mpi</td>
<td>2017-05-09 05:16:47</td>
<td>2017-05-09 05:21:06</td>
<td>977.1kB</td>
<td>1.1kB/op</td>
<td>499.8k</td>
</tr>
<tr>
<td>377680</td>
<td>1356</td>
<td>growfiles_mpi</td>
<td>2017-05-09 05:21:24</td>
<td>2017-05-09 05:26:54</td>
<td>984.1kB</td>
<td>1.7kB/op</td>
<td>399.6k</td>
</tr>
<tr>
<td>377695</td>
<td>1356</td>
<td>growfiles_mpi</td>
<td>2017-05-09 05:52:57</td>
<td>2017-05-09 05:57:21</td>
<td>986.4kB</td>
<td>1.1kB/op</td>
<td>409.3k</td>
</tr>
<tr>
<td>377688</td>
<td>1356</td>
<td>growfiles_mpi</td>
<td>2017-05-09 05:39:24</td>
<td>2017-05-09 05:43:45</td>
<td>987.2kB</td>
<td>1.7kB/op</td>
<td>447.8k</td>
</tr>
<tr>
<td>377673</td>
<td>1356</td>
<td>IOR</td>
<td>2017-05-09 05:13:55</td>
<td>2017-05-09 05:31:22</td>
<td>1.0MB</td>
<td>82.3MB/op</td>
<td>1.1k</td>
</tr>
<tr>
<td>377682</td>
<td>1356</td>
<td>IOR</td>
<td>2017-05-09 05:21:34</td>
<td>2017-05-09 05:39:02</td>
<td>1.0MB</td>
<td>90.5MB/op</td>
<td>1.1k</td>
</tr>
<tr>
<td>377690</td>
<td>1356</td>
<td>IOR</td>
<td>2017-05-09 05:43:51</td>
<td>2017-05-09 06:00:50</td>
<td>1.0MB</td>
<td>80.3MB/op</td>
<td>1.1k</td>
</tr>
<tr>
<td>377703</td>
<td>1356</td>
<td>IOR</td>
<td>2017-05-09 06:04:43</td>
<td>2017-05-09 06:27:43</td>
<td>1.0MB</td>
<td>116.5MB/op</td>
<td>1.1k</td>
</tr>
<tr>
<td>377696</td>
<td>1356</td>
<td>growfiles_mpi</td>
<td>2017-05-09 05:57:43</td>
<td>2017-05-09 06:02:07</td>
<td>1.0MB</td>
<td>1.7kB/op</td>
<td>130.3k</td>
</tr>
</tbody>
</table>

### Calculated Read Bandwidth
- Max: 2.95 GB/s
- Min: 0.0 MB/s

### Calculated Write Bandwidth
- Max: 0.0 MB/s
- Min: 0.0 MB/s
Design Patterns for Big Data /IOT Telemetry
InfluxDB Consumer

Kafka message bus

Infiniband provider

Syslogd

ElasticSearch

Syslog-ng

perfquery, ibnetdiscover

JobID Job Status

Jobevent provider

Lustre & Jobstats

Seastream provider

SNMP provider

SNX

Ingest

Integrate

Store / Access

Alerts

UI
Future Work
Caribou Future

- Caribou for Datawarp
  - LDMS collection joined to Caribou message bus
    - Diskstats
    - LVM
    - DWFS
    - XFS
    - DVS
- Lustre Client Stats?
- Interconnect?
Conclusion

- Big data/IOT style architecture for monitoring and metrics
- Differentiator for Sonexion
- Real-time streaming telemetry with customizable visualization
- Horizontal scalability across components
- Framework for future monitoring infrastructure