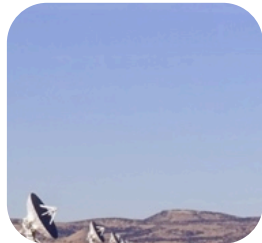
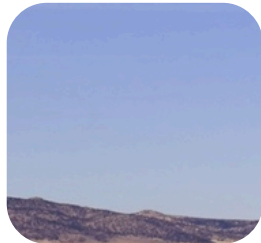


CRAY



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.

About Me



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 Region0ne

Tennat_id 529612f8fbc4332b66aab062afdf41

Component lustre

Device OST000a

device_type ost

Hostname unknown

Job_id 4058797

Product snx

Service storage

system_name snx11242

value 29461.066666666666

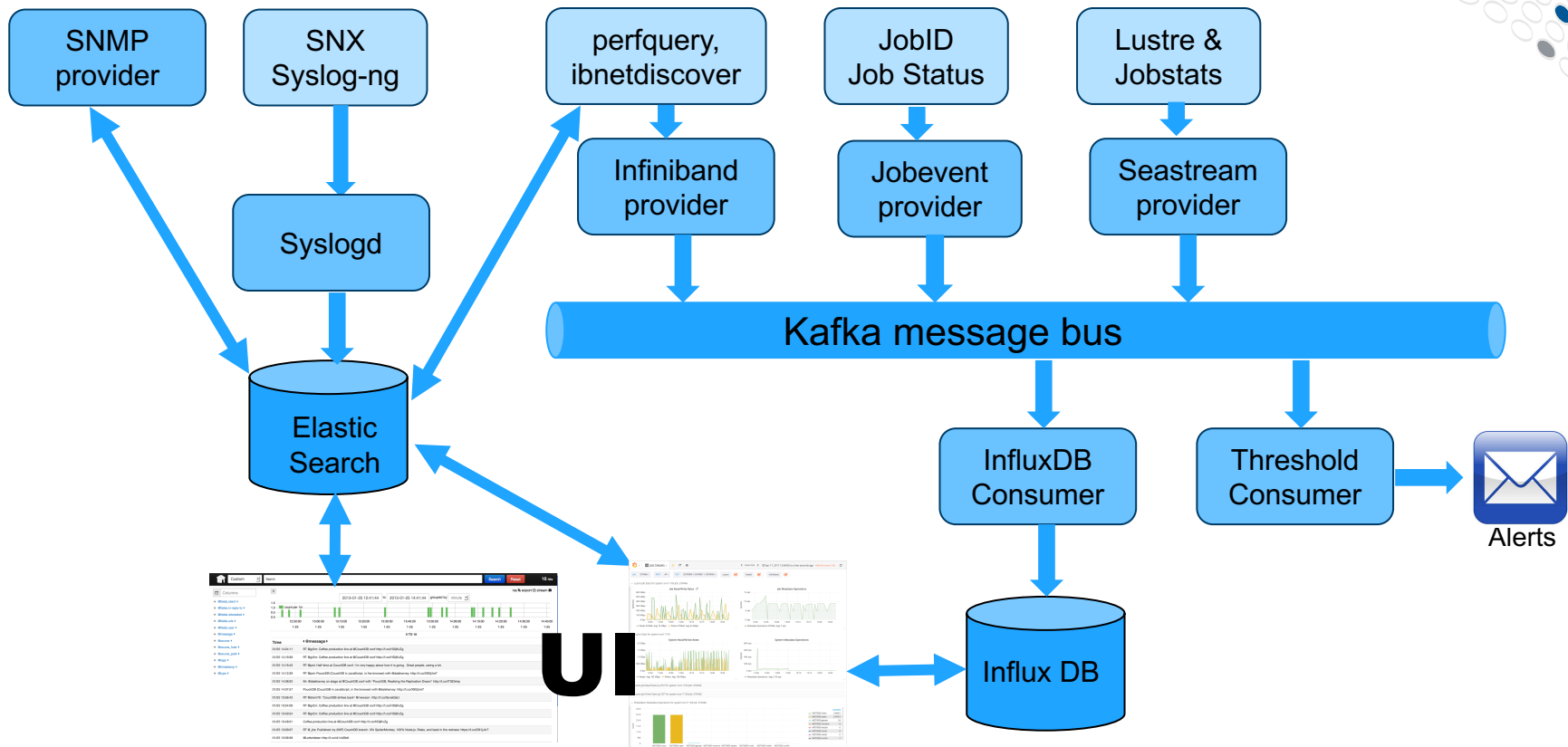
value_meta

Database Query Example

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

Time	_region system_name	_tenant_id value	component	device value_meta	device_type	hostname	job_idproduct	service	
2017-05-08T16:44:20Z 4058797	RegionOne snx	One storage	529612f8fbc4332b66aab062afdf41	snx11242	29461.066666666666	lustre	OST000a	ost	unknown
2017-05-08T16:44:21Z 4058797	RegionOne snx	One storage	529612f8fbc4332b66aab062afdf41	snx11242	48251.2	lustre	OST0008	ost	unknown
2017-05-08T16:44:24Z 4058797	RegionOne snx	One storage	529612f8fbc4332b66aab062afdf41	snx11242	98330.666666666667	lustre	OST0000	ost	unknown
2017-05-08T16:44:26Z 4058797	RegionOne snx	One storage	529612f8fbc4332b66aab062afdf41	snx11242	68594.4	lustre	OST0003	ost	unknown

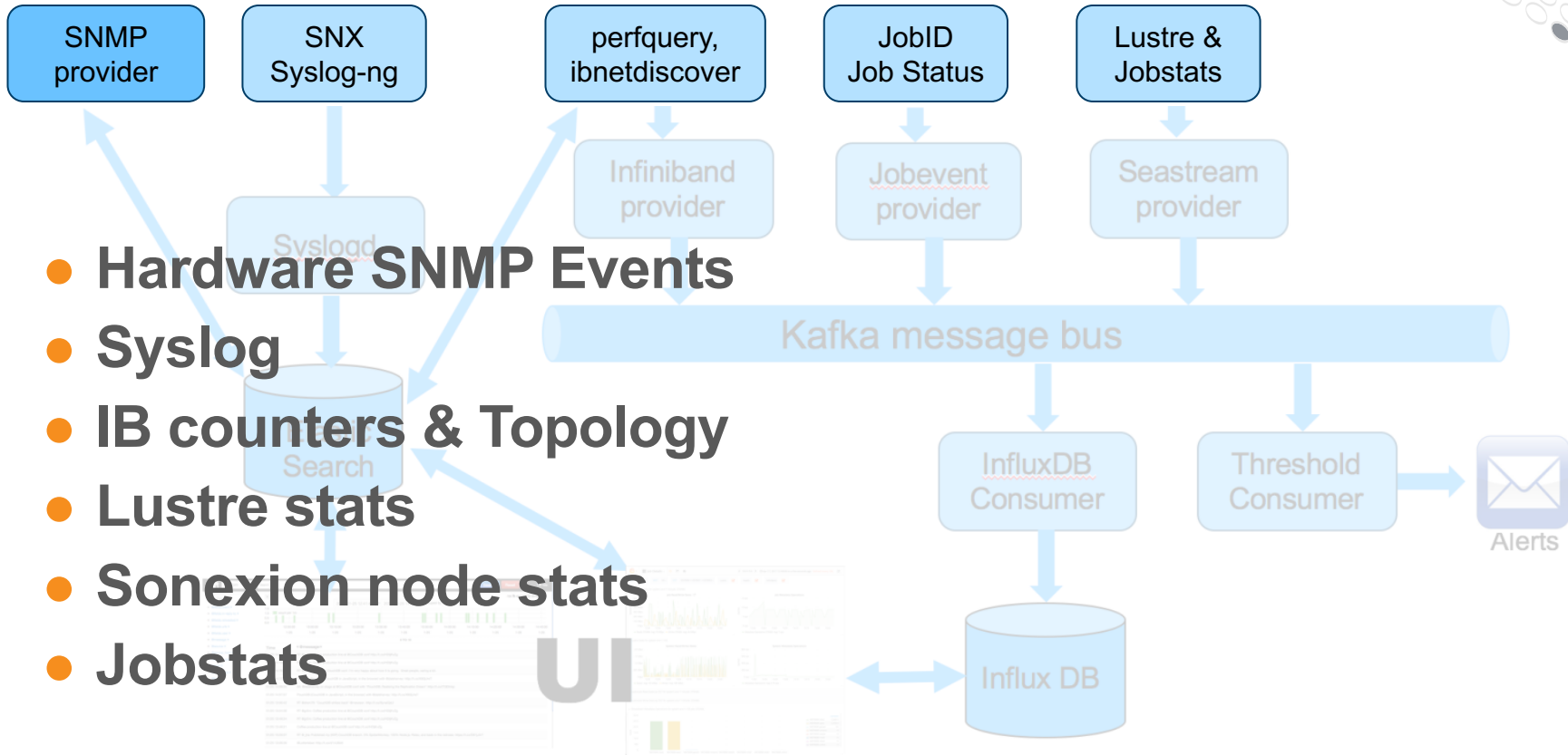
Data Collection



COMPUTE

STORE

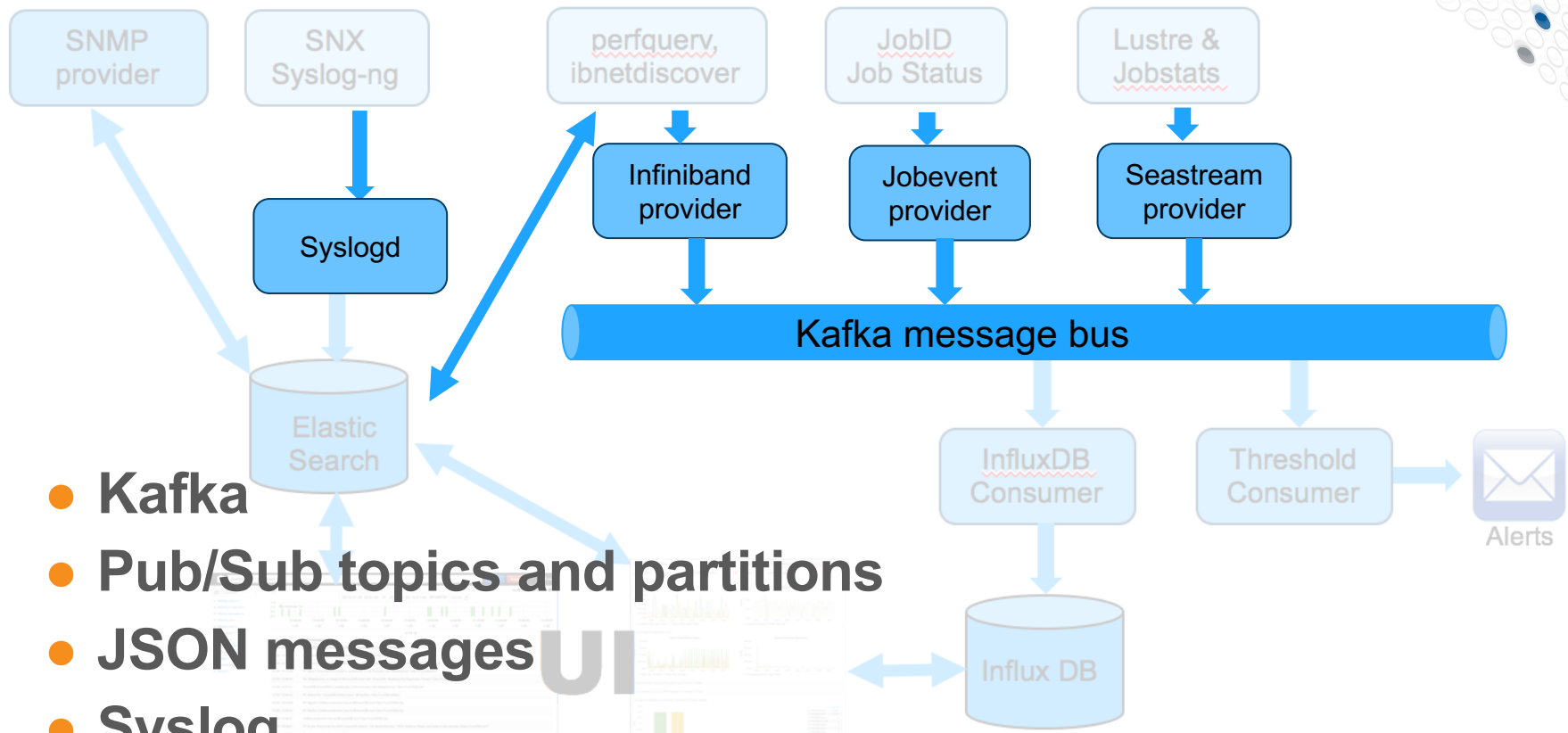
ANALYZE



- **Hardware SNMP Events**
- **Syslog**
- **IB counters & Topology**
- **Lustre stats**
- **Sonexion node stats**
- **Jobstats**

COMPUTE | STORE | ANALYZE

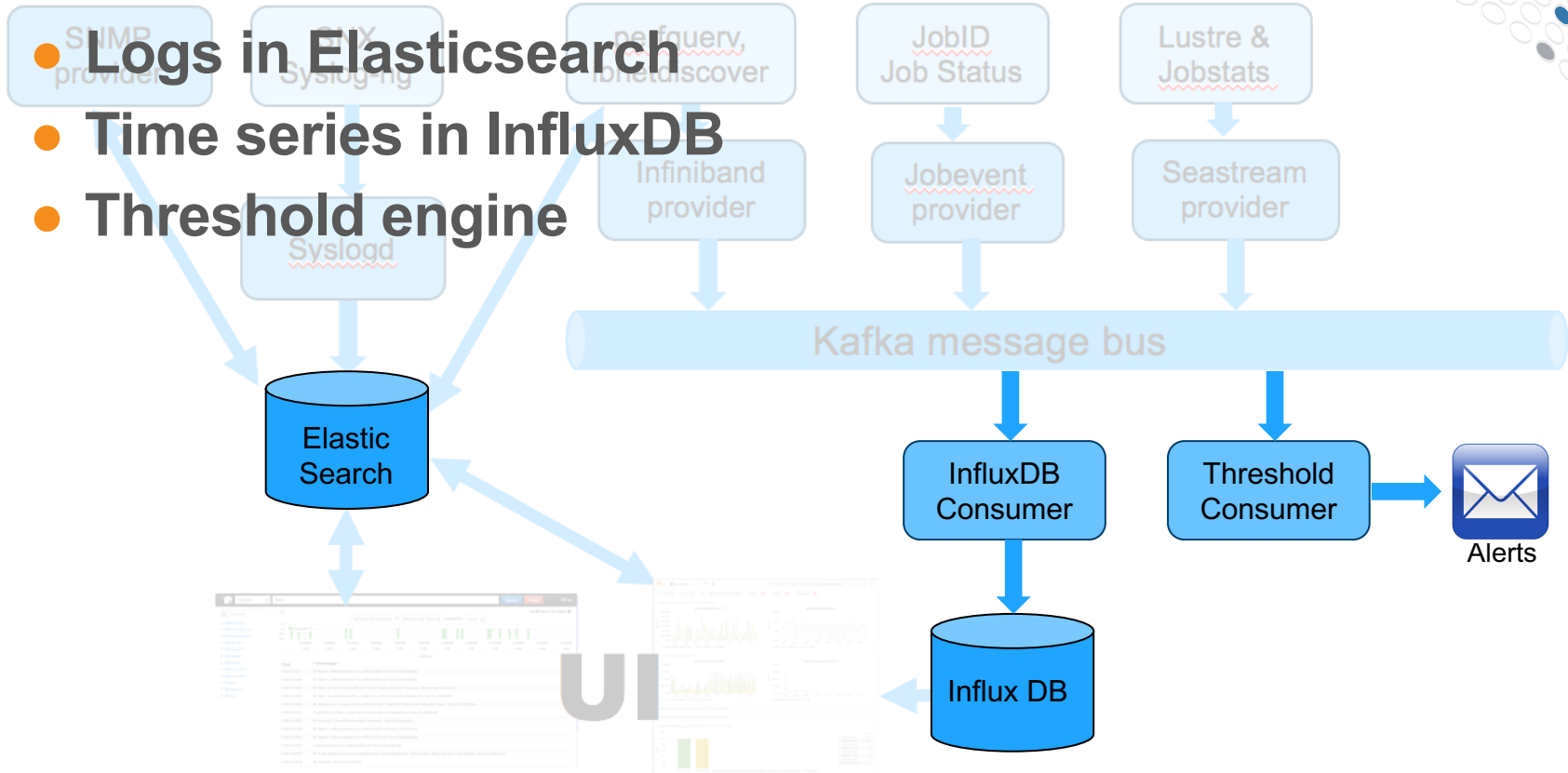
Data Integration and Transport



- **Kafka**
- **Pub/Sub topics and partitions**
- **JSON messages**
- **Syslog**

Data Persistence

- Kafka consumers
- Logs in Elasticsearch
- Time series in InfluxDB
- Threshold engine



COMPUTE

STORE

ANALYZE

User Interface and Workflows

- Identity ▾
- Alarms and Notifications ▾
- Statistics ▲

Caribou

Last 15 minutes

Sonexion Status

snx11103
✖ ✎

OST I/O	Metadata ops	Capacity
1.73 GB/s Average Read	0.01 k/s Requests	71.26 % Available
672.44 MB/s Average Write	451.61 TB Total	
3 / 0 total / warning	11 / 15 switches / HCAs	4 / 0 total / warning

No Health Events

snx11242
✖ ✎

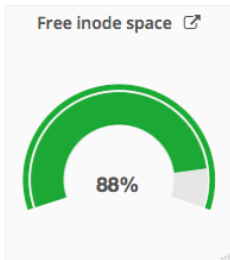
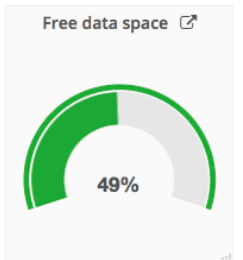
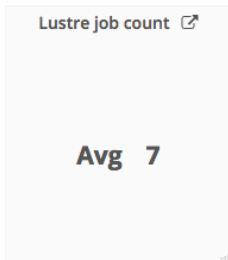
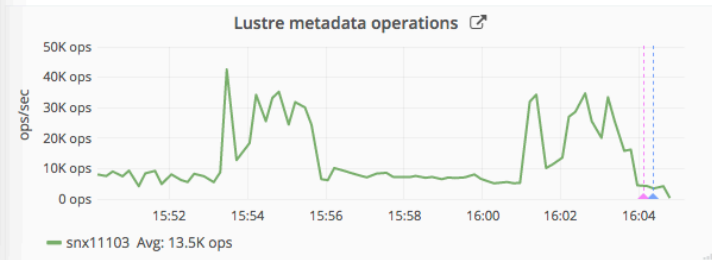
OST I/O	Metadata ops	Capacity
1.60 kB/s Average Read	0.00 k/s Requests	53.28 % Available
30.40 MB/s Average Write	1.32 PB Total	
0 / 0 total / warning	0 / 0 switches / HCAs	12 / 0 total / warning

No Health Events

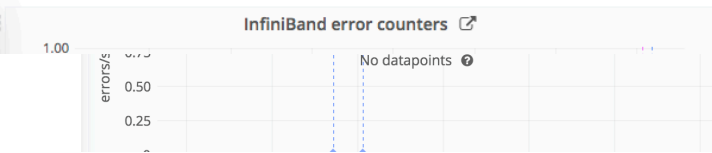
Add Sonexion



▼ Lustrre metrics



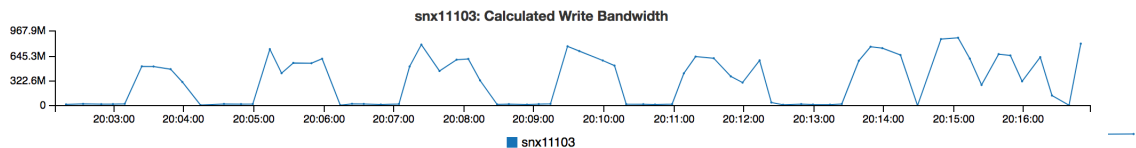
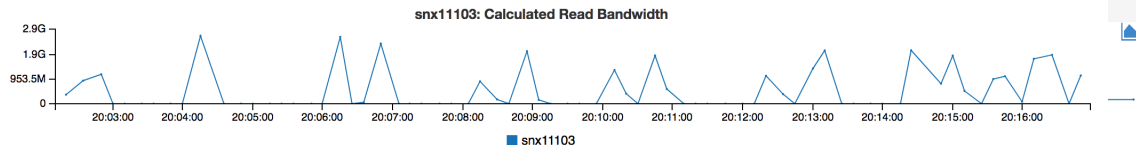
▼ InfiniBand metrics

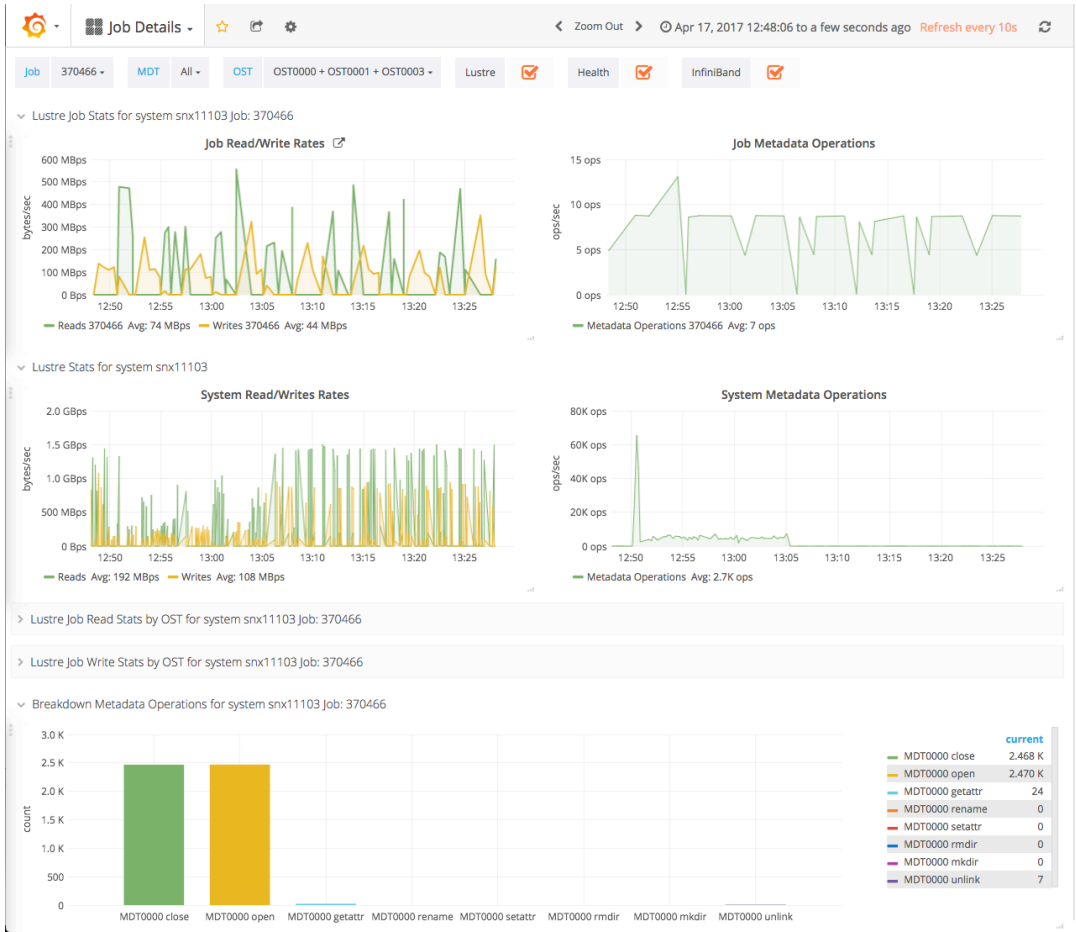


Jobs

- Identity
- Alarms and Notifications
- Statistics
- Sonexion Status

377678	1356	growfiles_mpi	2017-05-09 05:16:49	2017-05-09 05:21:16	963.0kB	1.2kB/op	497.1k
377700	1356	growfiles_mpi	2017-05-09 06:01:23	2017-05-09 06:05:41	970.5kB	1.6kB/op	501.6k
377677	1356	growfiles_mpi	2017-05-09 05:16:47	2017-05-09 05:21:06	977.1kB	1.1kB/op	499.8k
377680	1356	growfiles_mpi	2017-05-09 05:21:24	2017-05-09 05:26:54	984.1kB	1.7kB/op	399.6k
377695	1356	growfiles_mpi	2017-05-09 05:52:57	2017-05-09 05:57:21	986.4kB	1.1kB/op	409.3k
377688	1356	growfiles_mpi	2017-05-09 05:39:24	2017-05-09 05:43:45	987.2kB	1.7kB/op	447.8k
377673	1356	IOR	2017-05-09 05:13:55	2017-05-09 05:31:22	1.0MB	82.3MB/op	1.1k
377682	1356	IOR	2017-05-09 05:21:34	2017-05-09 05:39:02	1.0MB	90.5MB/op	144.0
377690	1356	IOR	2017-05-09 05:43:51	2017-05-09 06:00:50	1.0MB	80.3MB/op	409.0
377703	1356	IOR	2017-05-09 06:04:43		1.0MB	116.5MB/op	145.0
377696	1356	growfiles_mpi	2017-05-09 05:57:43	2017-05-09 06:02:07	1.0MB	1.7kB/op	130.3k





COMPUTE

STORE

ANALYZE

- Identity ▾
- Alarms and Notifications ▲

Alarm Definitions

Filter

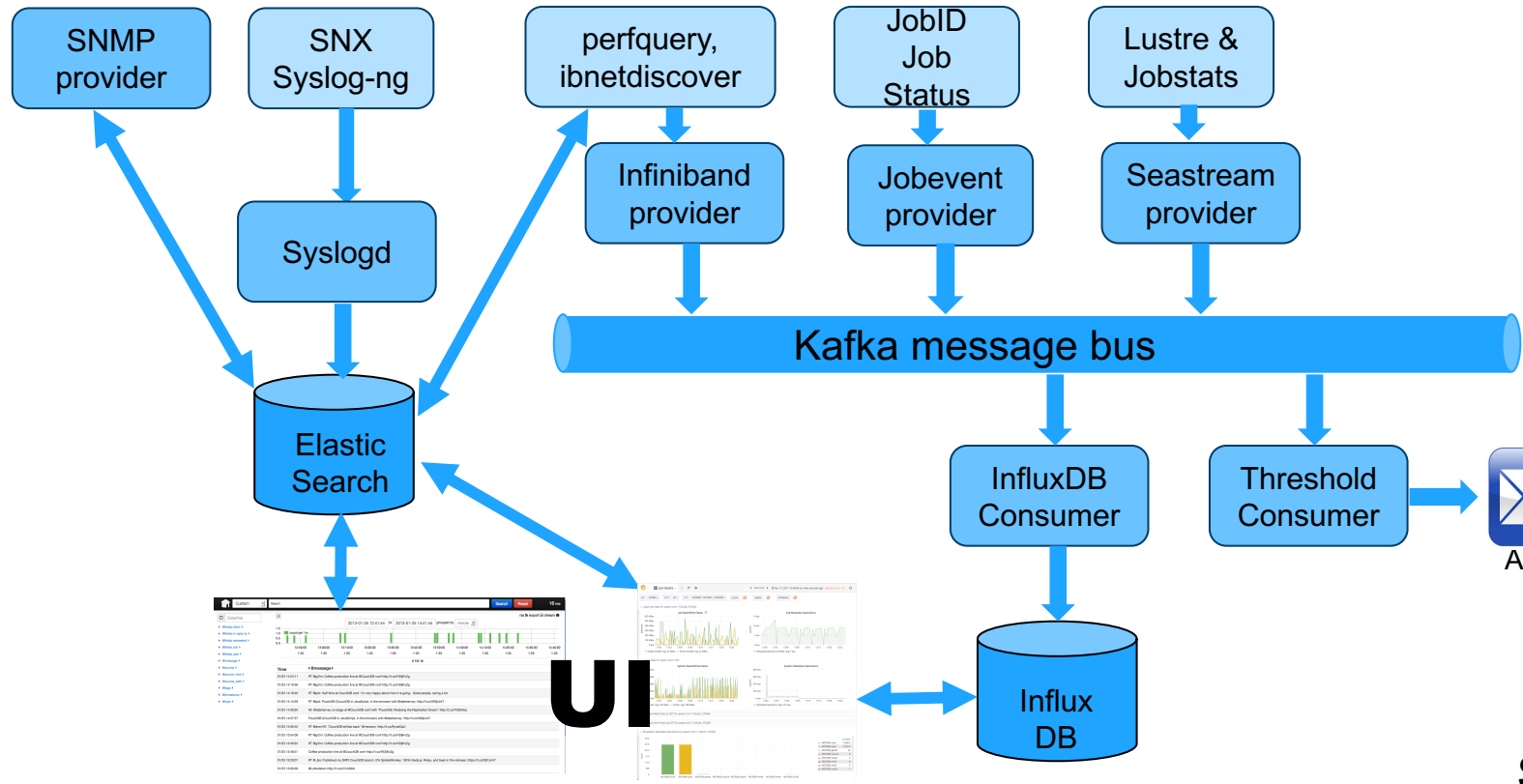
Alarm Definitions

- Alarms
- Notifications
- Statistics ▾

<input type="checkbox"/>	Name	Description	Notifications Enabled	Actions
<input type="checkbox"/>	Caribou_Sonexion_snmpwalk_timeout	There is a timeout using SNMPwalk to a Sonexion	True	<input type="button" value="Edit Alarm Definition"/> ▾
<input type="checkbox"/>	Caribou_IB_Topology	IB Topology Change Detection	True	<input type="button" value="Edit Alarm Definition"/> ▾
<input type="checkbox"/>	Var_Filesystem_Disk_Space_Usage	The disk space of the /var filesystem partition is over 95%	True	<input type="button" value="Edit Alarm Definition"/> ▾
<input type="checkbox"/>	Caribou_OST_Free_Space	The free disk space of a specific OST is under 5%	True	<input type="button" value="Edit Alarm Definition"/> ▾
<input type="checkbox"/>	Root_Partition_inode_Usage	The use of the configured inodes of the root partition is over 95%	True	<input type="button" value="Edit Alarm Definition"/> ▾
<input type="checkbox"/>	cray_ib.symbol_errors_sec	IB SymbolErrors exceeds 120/s	True	<input type="button" value="Edit Alarm Definition"/> ▾
<input type="checkbox"/>	Caribou_OST_Free_Files	The free inode space of a specific OST/MDT is under 5%	True	<input type="button" value="Edit Alarm Definition"/> ▾
<input type="checkbox"/>	Caribou_Sonexion_Metric_Health	Check for Incoming metrics	True	<input type="button" value="Edit Alarm Definition"/> ▾
<input type="checkbox"/>	Keystone_Services_are_down	All the Keystone services are not operating properly	True	<input type="button" value="Edit Alarm Definition"/> ▾
<input type="checkbox"/>	Root_Partition_Disk_Space_Usage	The disk space of the root partition is over 95%	True	<input type="button" value="Edit Alarm Definition"/> ▾

Displaying 10 items

Design Patterns for Big Data /IOT Telemetry



←
Ingest

←
Integrate

←
Store / Access

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



Q&A

Craig Flaskerud
cflaskerud@cray.com

CUG.2017.CAFFEINATED COMPUTING

Redmond, Washington May 7-11, 2017