Hewlett Packard Enterprise

CLUSTER HEALTH CHECK DIAGNOSTICS SUITE

Amarnath Chilumukuru - amarnath.c@hpe.com
Prasanth Kurian - prasanth.kurian@hpe.com

January 31, 2022

CLUSTER HEALTH MANAGEMENT CAPABILITIES

Cluster Health Management addresses data center concerns:

Datacenter Director

- Cluster is down, I'm up!
- Cluster delivers maximum performance and up-time
- Time is money

System Admin

- Needs to get ahead of the problem—detect hardware issues before they turn into failures
- The job queue is over-capacity, can't afford downtime or slowness
- Need to maintain planned maintenance schedule—no unscheduled downtime

Researcher: Data Scientist, Data Analyst, Engineer…

- Why is my job running slow?
- Why did my job fail?
- Getting ready to launch a long-running job, is the cluster healthy?

Cluster Health Management capabilities:

System Administrators

- **Minutes:** On-demand cluster health checks, self-service diagnostics
- **Hours:** Invasive health tests run during planned maintenance window or for triage from failures (same as the Factory Compliance Tests)
- **Real-Time** Manage automated alerts on failures or pre-defined actions from Cluster Health Dashboard

Users

- Resource management tool does a quick job-level health check (seconds)
- Supports Slurm and Altair PBS Professional to validate compute nodes before launching jobs

HPCM CLUSTER HEALTH CHECK & CSM DIAGNOSTICS

Portfolio

System Admin

- Health checks on Periodic Maintenance Cycle to identify the potential problems
- Health checks integrated with Job schedulers to ensure smooth job runs on healthy hardware
- Pre-Flight health checks before hardware being put into production
- FRU level Diagnostics to triage the hardware issues for easy identification of problems

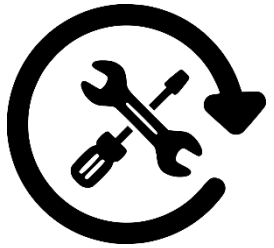
Researcher

Coverage

Hardware

All the HPE HPC hardware platforms like Cray EX, Apollo 9000, 2000, 6500, 6000, 70, 35, 20, ProLiant DL servers and SGI 8600.

HPE Cluster Manager (HPCM) and Shasta



Cluster Manager

SHORTER TIME TO PRODUCTION WITH SYSTEM DIAGNOSTICS

HPE Factory

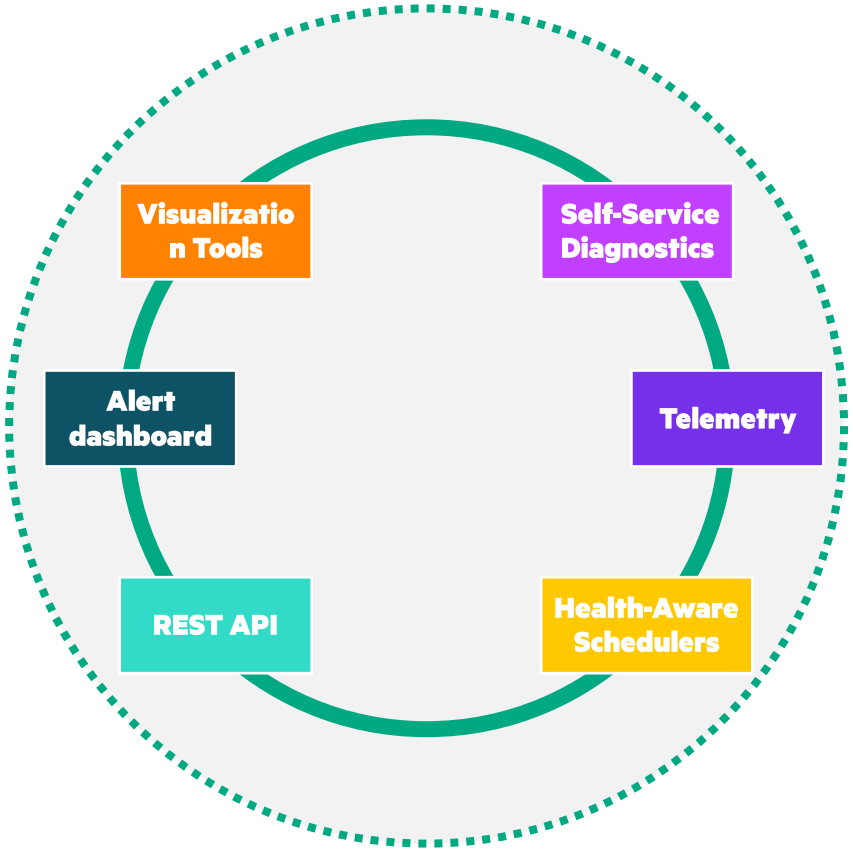
New System Setup

We supply the same system diagnostics used by HPE factory to customers for compliance check

Customer

Test	Description	Test	Description
CPU Uniformity Checks	Finds the outlier nodes with non-uniform configuration	**Check the Health of the InfiniBand Fabric**	Link integrity checks, Missing links, Down links, link speed and width, Subnet manager availability etc.
CPU Performance Check	Runs High Performance Linpack on each compute nodes and compares the GFLOPS reported on each node by High-Performance LINPACK (HPL).	**Check the Health of Gluster Filesystem**	Check the Peer node availability, Availability of Volumes, Availability of bricks etc.
Cluster-wide CPU Performance Check	Runs the cluster-wide High Performance Linpack on a cluster.	**Disk Health Check**	Writes different patterns to disk and read the data back to check the consistency.
Memory Uniformity Checks	Finds the outlier nodes with non-uniform configuration.	**I/O Performance Check Tool**	Used to benchmark disk performance. Detects the low performing disks on each node.
Memory Performance Check	Runs the STREAM benchmark on each node and reports the TRIAD value. This check reports low performing nodes.	**Filesystem Performance Check**	Benchmarking and in I/O performance measurement. Detects the slow filesystems and helps to identify the bottlenecks for filesystem finetuning.
Ethernet Uniformity Checks	Finds the outlier nodes with non-uniform fabric configuration.	**GPU Health Check Tool**	Stress the GPUs to find out the Errors and identify faulty GPUs.
Fabric Checks	Health of Mellanox InfiniBand HCAs and HPE Slingshot high-speed network (HSN) Cassini cards.	**CDU Checks**	Reports the health of the cooling distribution units (CDUs) in the cluster.
Fabric Performance Health Check	Conducts a point-to-point MPI bandwidth test to measure InfiniBand fabric performance. Highlights the outlier nodes.	**Console Checks**	Check the console availability of compute nodes. Report the nodes with inaccessible consoles.

CLUSTER HEALTH CHECK DIAGNOSTICS - KEY FEATURES



Diagnostics

Self-Service diagnostics for easy Cluster Service and Maintenance
Light weight and Interactive Reports
Health checks for quick inspection in production environment
Stress tests during planned maintenance window
Customizable Test Plans
Coverage of major cluster components

Cluster Monitoring & Telemetry

Visualization dashboards for the cluster health and performance metrics
Easy & Powerful Console based and graphical dashboards
REST APIs for third party software integrations
Real time Email alerts

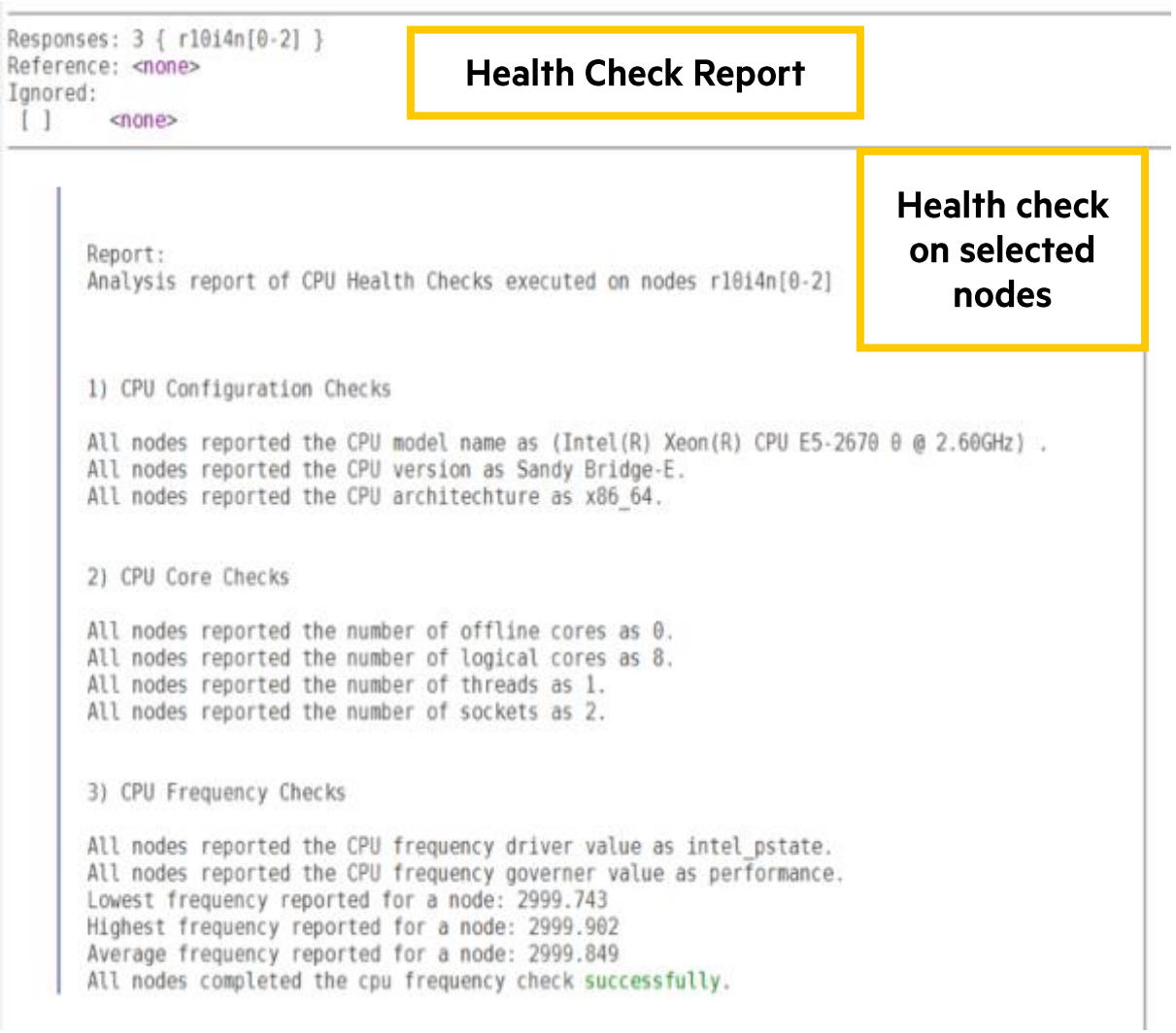
Health-Aware Job Schedulers

Prolog/Epilog and periodic node health monitoring
Hooks for PBS and SLURM
Ability to validate compute nodes prior to application launch

ON-DEMAND CLUSTER HEALTH CHECK

- Request health check anytime with data reported in a few seconds
- Hardware level diagnostics for node, chassis, rack, cpus, memory, disks, i/o, Lustre
- Fabric diagnostics
- Firmware/software versions, configuration settings

```
[root@iceadmin ~]# cm health -l
      cpuchk
      cpuperf
      cwcpuperf
      memchk
      memperf
      netchk
      fabricchk
      fabricperf
[root@iceadmin ~]#
```



CLUSTER HEALTH DIAGNOSTICS AVAILABLE

Key Hardware

CPU
- Model, Architecture, Online vs Offline Cores, threads
- Industry Standard benchmarks to check CPU Performance

Memory
- Available Memory, Swap space, THP, DIMM Size, DIMM Speed, Balanced DIMM Configurations
- Industry standard benchmarks for Memory Performance Checks

Ethernet
- Link Speed, Link Detection, Drivers, Firmware, PCI slots, pings

Key System Components
- PCI speed, Firmware , Fans, CDU

Disks
- Errors, Bad blocks,
- IOPS, Bandwidth, latency

GPU
- Identify Faulty GPUs

Key Software

Schedulers
- PBS and SLURM
- Node Health and Job Health

GlusterFS
- Node, Volume, Brick Health
- Warns on bricks usage limits

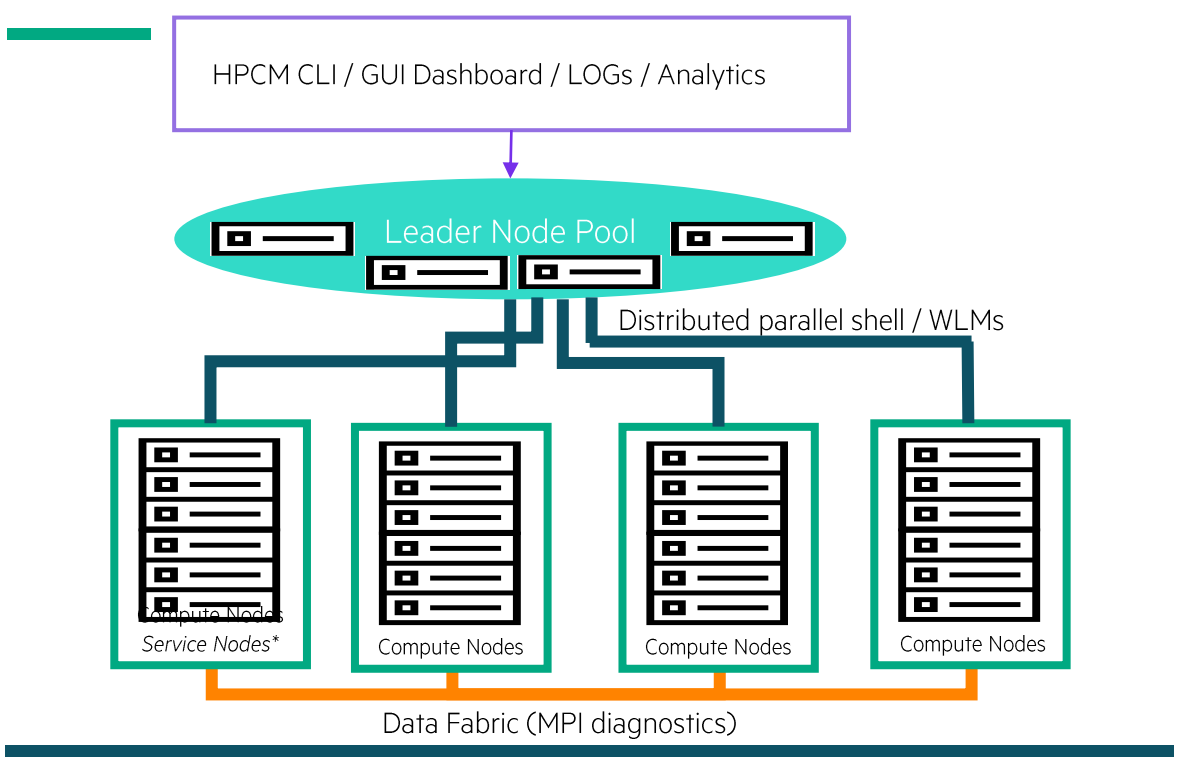
Filesystems
- Bad blocks, Filesystem Errors
- Industry standard Performance Benchmarks

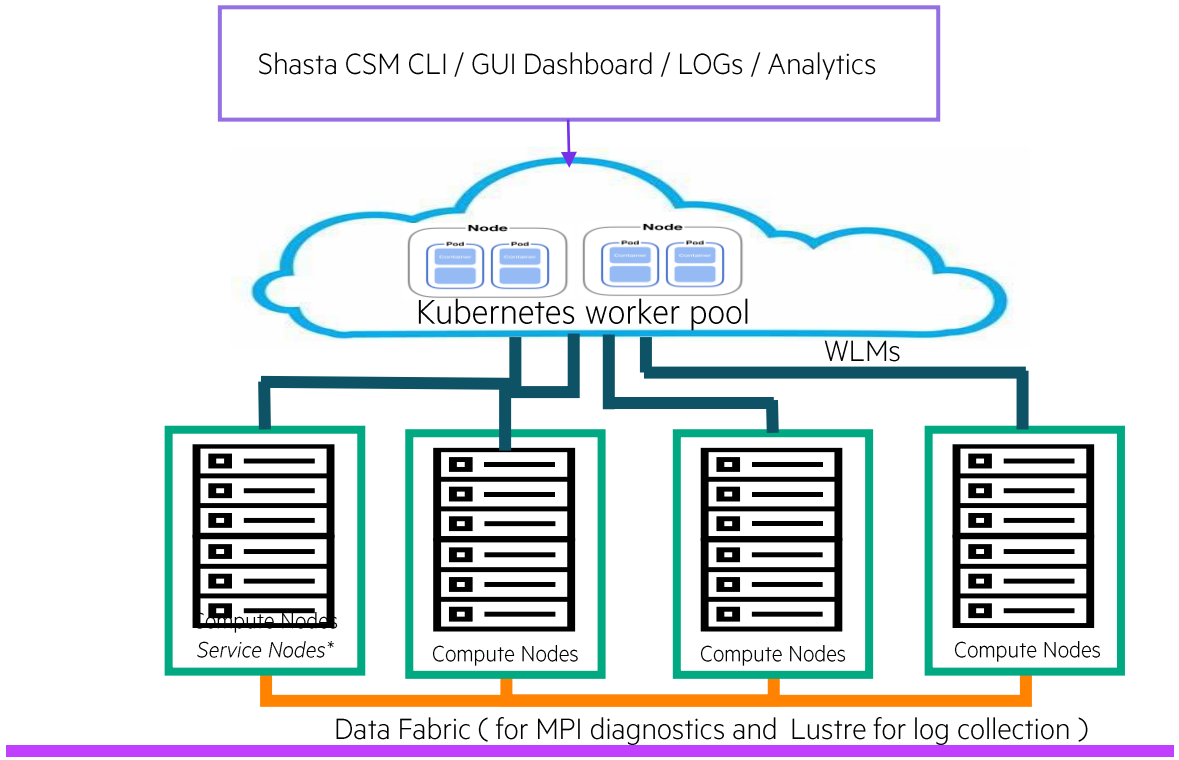
Interconnects

InfiniBand Fabric
- Host Level and Topology Level
- HCA Type, Speed & width, NUMA consistency
- Link Speed, Link width, Missing Links, Link Integrity
- Standard Benchmarks for Fabric Performance Checks

Slingshot Fabric
- Host Level and Topology Level
- HCA Type, Speed & width, NUMA consistency
- Link Speed, Link width
- Missing Links, Link Integrity
- Standard Benchmarks for Fabric Performance Checks

ON DEMAND DIAGNOSTICS SCALABLE ARCHITECTURE





- Diagnostics can run using WLMs (SLURM and PBS)
- Logs are collected at central shared location
- Analysis on logs and Summarized Reports
- Ability to customize the diagnostics and to launch a batch of diagnostics

CLUSTER HEALTH DASHBOARD

At-a-glance cluster health overview and links to more detailed reports

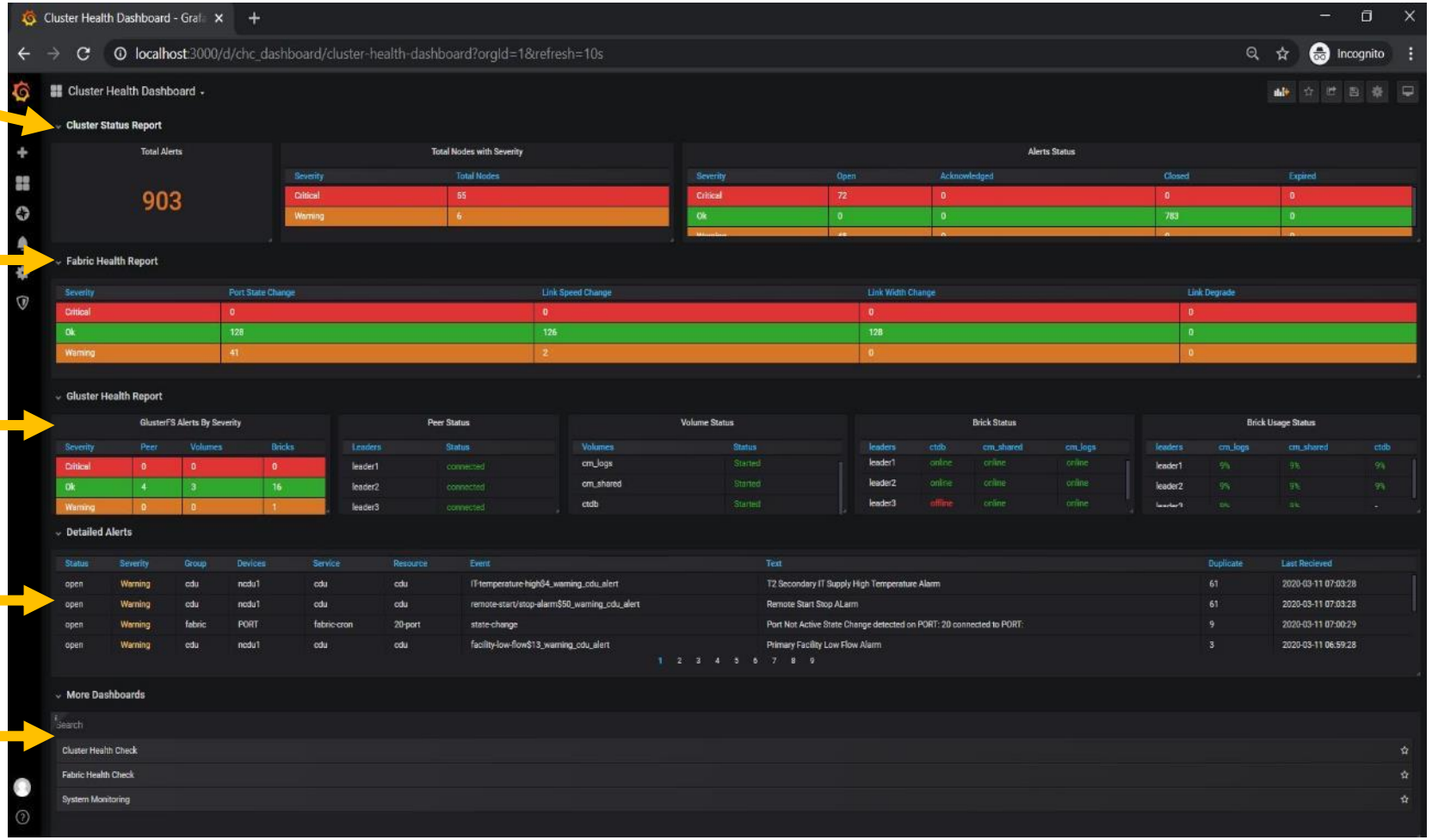
Cluster Status Report
Alert numbers as well as information on nodes with severity and alerts status

Fabric Health Report
Includes fabric port-state changes, link speed changes, link width changes, and link degrade changes.

Cluster Health Report
Includes severity levels and includes peer, volume, brick, and brick usage status.

Detailed Alerts
Complete details of the various alerts.

More Dashboards
Link to other dashboards like System Monitoring, Cluster Health Check, Fabric Health Check, CDU monitoring and Scheduler Status Report.

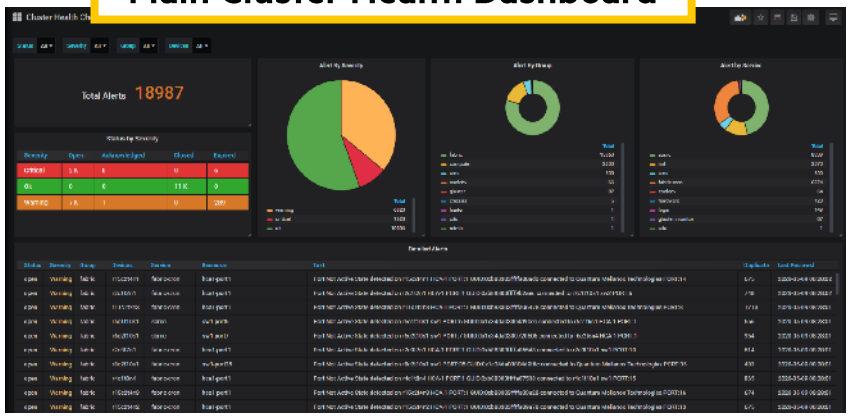


REAL-TIME CLUSTER HEALTH MONITORING

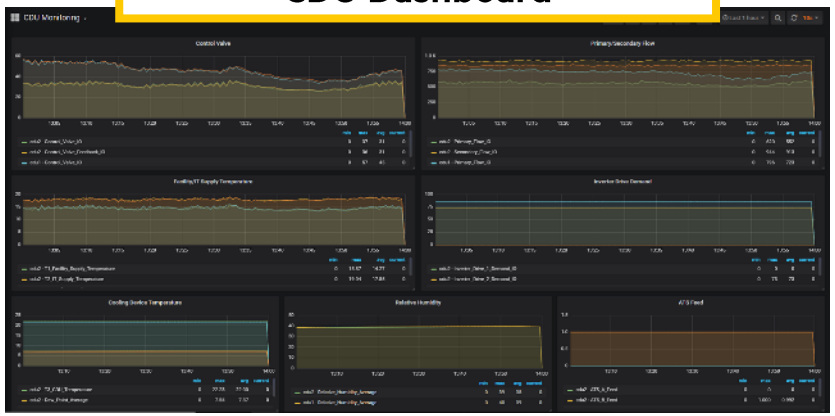
Visualization dashboards

- **Cluster Health at-a-Glance—** Single Pane view for the complete cluster Health Status
- **Live System Monitoring—** Dashboards for key metrics like Power, Cooling, CPU, Memory, Disk, Fabric, Gluster, Job Scheduler monitoring metrics
- **Scalable—**Highly scalable data pipeline at the backend
- **Customizable**—Create new dashboards easily

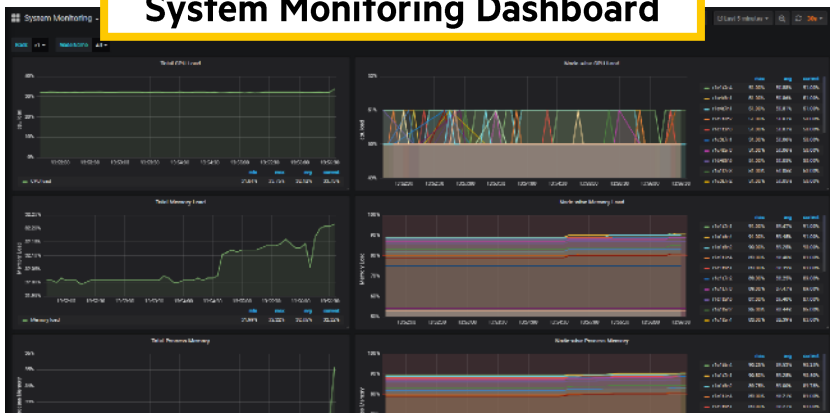
Main Cluster Health Dashboard



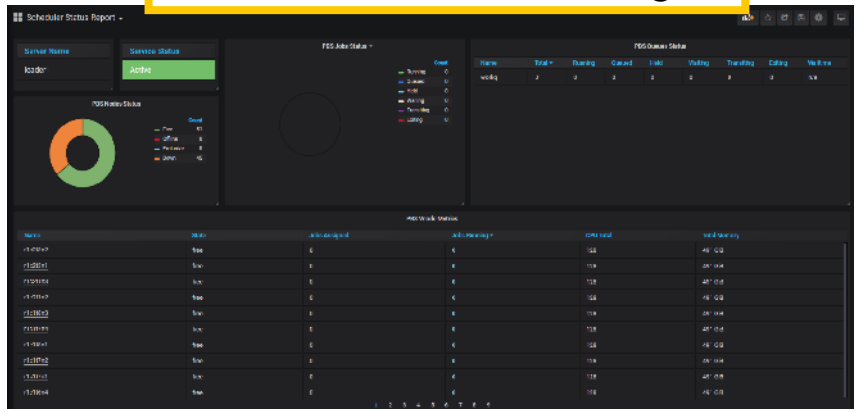
CDU Dashboard



System Monitoring Dashboard



Jobs Scheduler Monitoring



CLUSTER HEALTH ALERTING DASHBOARD

- Cluster Alert management dashboard
- Configure & manage system telemetry and alerts
- Includes ability to analyze alerts on historical system events
- Covers iLO / system log files, FRU, CDU, Fabric, CMC, Hardware events, Job Scheduler monitoring, Gluster, Performance metrics

Current alerts status “top”

```
[root@himgiri ~]# cm health alert top
http://localhost:9090

 Sev. Time      Dupl. Customer Env.          Service      Resource     Group Event         Value Text
 Ok   12:05:32     0 -        iceadmin      DIMM         dimm         admin dimm_clock_s dimm  iceadmin:dimm speed same
 Warn 10:58:12     0 -        r10lead       blades       blade        leade blade_replac new m Rack r1: blade replaced
 Warn 18:22:09     0 -        r1i3n37       blades       blade        ice_c new_blade_de new h Rack r1: new blade detected
 Warn 18:20:59     0 -        r1i2n18       blades       blade        ice_c new_hostname new h Rack r1: new host detected
 Warn 18:20:52     0 -        r1i2n17       blades       blade        ice_c new_hostname new h Rack r1: new host detected
 Warn 18:20:36     0 -        r1i2n16       blades       blade        ice_c new_hostname new h Rack r1: new host detected
 Crit 16:31:41     0 -        r1i3n11       blades       blade        ice_c blade_pulled blade Rack r1: blade pulled out
```

Query alerts on specific blade

```
[root@himgiri ~]# cm health alert query -i 63675ca2
STATUS    TEXT                          LAST RECEIVED        GROUP         DUPL  ID        CUSTOMER    RESOURCE    SEVERITY    SERVICE    VALUE                    EV
 ENVIRONMENT
-------- ---------------------------- ------------------- ----------- ------ -------- ---------- ---------- ---------- --------- ---------------------- --
 -------------
open      Rack r1: new blade detected  2019/10/16 18:22:09  ice_compute       0  63675ca2            blade       warning     blades     new hostname and mac id  ne
 r1i3n37
[root@himgiri ~]#
[root@himgiri ~]#
[root@himgiri ~]# cm health alert query -f severity=warning
STATUS    TEXT                          LAST RECEIVED        GROUP         DUPL  ID        CUSTOMER    RESOURCE    SEVERITY    SERVICE    VALUE
           ENVIRONMENT
-------- ---------------------------- ------------------- ----------- ------ -------- ---------- ---------- ---------- --------- ------------------------------ --------
------------ -------------
expired   Rack r1: blade replaced      2019/10/17 10:58:12  leader            0  e15a8c24            blade       warning     blades     new mac id but hostname unchanged  blade_rep
laced         r10lead
open      Rack r1: new blade detected  2019/10/16 18:22:09  ice_compute       0  63675ca2            blade       warning     blades     new hostname and mac id            new_blade
_detected     r1i3n37
open      Rack r1: new host detected   2019/10/16 18:20:59  ice_compute       0  b5d7c305            blade       warning     blades     new hostname but unchanged macid   new_hostn
ame_detected  r1i2n18
open      Rack r1: new host detected   2019/10/16 18:20:52  ice_compute       0  0da40acd            blade       warning     blades     new hostname but unchanged macid   new_hostn
ame_detected  r1i2n17
open      Rack r1: new host detected   2019/10/16 18:20:36  ice_compute       0  181aaca9            blade       warning     blades     new hostname but unchanged macid   new_hostn
ame_detected  r1i2n16
```

Summary of alert activity across the cluster

```
[root@himgiri ~]# cm health alert -s

Alert Status        Count
------------        -----
Critical            1
Warnings            5
Information         1


Open                5
Acknowledged        0
Closed              1
Expired             1


Group               Severity        Alerts
------              --------        ------
admin               ok              critical : 0, warning : 0, info : 1
cmc                 ok              critical : 0, warning : 0, info : 0
compute             ok              critical : 0, warning : 0, info : 0
ib_switch           ok              critical : 0, warning : 0, info : 0
ice_compute         critical        critical : 1, warning : 4, info : 0
leader              warning         critical : 0, warning : 1, info : 0
leader_alias        ok              critical : 0, warning : 0, info : 0
mgmt_switch         ok              critical : 0, warning : 0, info : 0
pdu                 ok              critical : 0, warning : 0, info : 0
switch_blade        ok              critical : 0, warning : 0, info : 0

[root@himgiri ~]#
[root@himgiri ~]# cm health alert ice_compute

ice_compute         Severity        Summary
-----------         ---------       -------
r1i2n18             warning         new_hostname_detected:1
r1i3n11             critical        blade_pulled_out:1
r1i2n16             warning         new_hostname_detected:1
r1i3n37             warning         new_blade_detected:1
r1i2n17             warning         new_hostname_detected:1
```

SEND ALERTS IN EMAIL

- Elastalert sends the alerts in email as they are generated.
- Configuration file is available in /opt/clmgr/elastalert/mail_config.yaml
- By default, these alerts are disabled.

Email Message Content

Current Open Alerts in the Cluster

STATUS	SEVERITY	GROUP	ENVIRONMENT	SERVICE	RESOURCE	EVENT	VALUE	TEXT	DUPLICATES	LAST RECEIVED
open	critical	cdu	node178	test_service	test_resource	cdu_event1	test_value	Testing the elastalert email rule	0	2020-03-25 03:48:29 CDT
open	warning	gluster	node176	test_service	test_resource	test_event2	test_value	Testing the elastalert email rule	1	2020-03-25 03:48:33 CDT
open	critical	cdu	node184	test_service	test_resource	cdu_event2	test_value	Testing the elastalert email rule	0	2020-03-25 03:48:35 CDT
open	critical	fabric	node58	test_service	test_resource	test_event	test_value	Testing the elastalert email rule	0	2020-03-25 03:48:38 CDT
open	critical	leader	leader1	test_service	test_resource	leader1_event	test_value	Testing the elastalert email rule	1	2020-03-25 03:48:42 CDT

HEALTH-AWARE JOB SCHEDULERS

Avoid launching jobs on unhealthy nodes

Reliability

- Prevent Job failures due to hardware and software problems

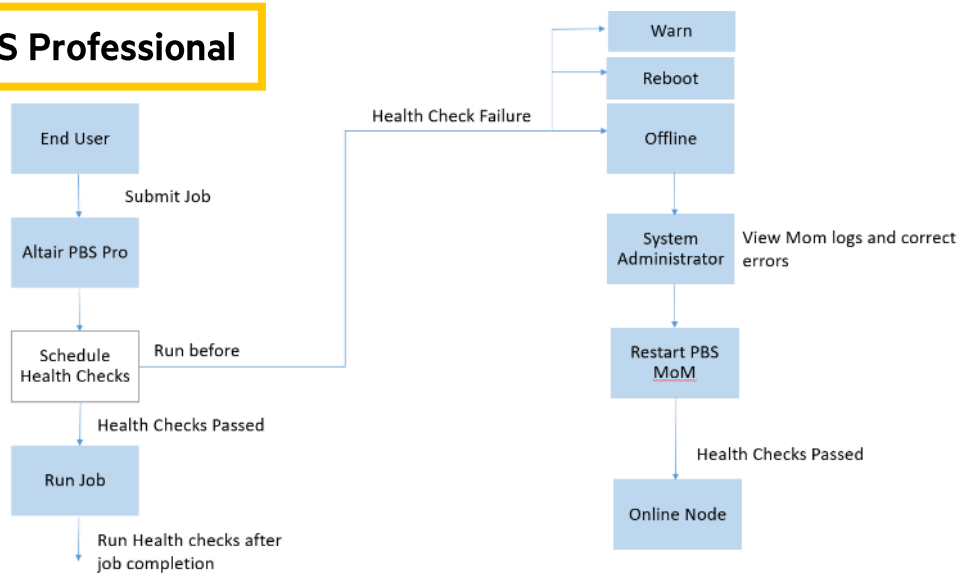
Reduced maintenance time

- Automated detection and notification of unhealthy nodes

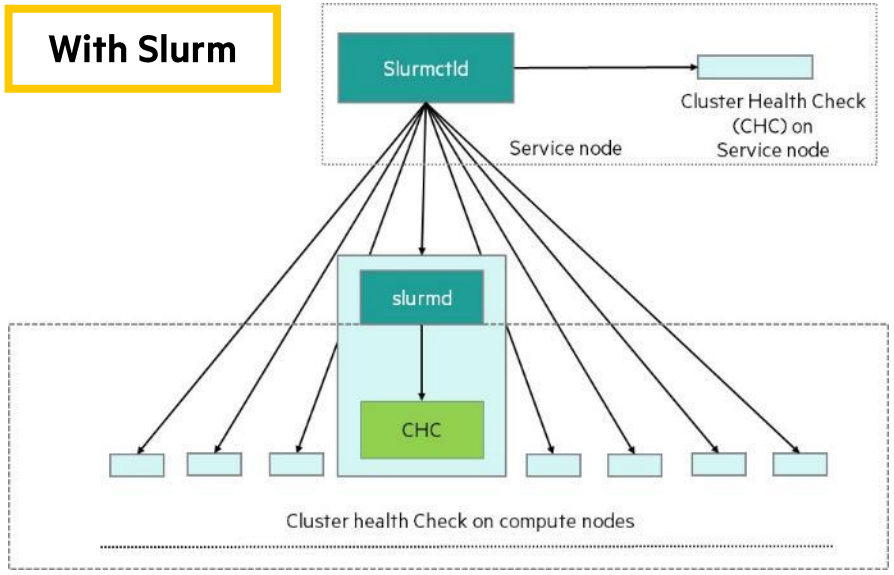
Customizable runs

- Periodic health checks
- Prologue & Epilogue health checks

with PBS Professional



With Slurm



THANK YOU