Integrated monitoring of the ATLAS online computing farm

S. Ballestrero · F. Brasolin · D. Fazio · C. Gament · C.J. Lee · D.A. Scannicchio · M.S. Twomey

Why

The online farm of the ATLAS experiment at the LHC consists of nearly 4000 PCs with various characteristics and the status and health of every host must be constantly monitored to ensure the correct and reliable operation of the whole online system.

The monitoring is the first line of defense: it should not only promptly provide alerts in case of failure but, whenever possible, warn of impending issues.

How

Two different monitoring systems have been combined to get a complete picture of the system:

- ★ Icinga 2 for active checks and alerting
  - + it replaced Icinga (end 2015) which in turn replaced Nagios v3 (2014/2015)
  - + it improved scalability and has native support for load balancing
    - ◗ it currently handles ~ 4000 hosts and ~ 83200 checks performed with different time intervals ranging between 5 minutes and 24 hours according to the need
  - + two servers are configured with the built-in High Availability Cluster feature
- ★ Ganglia for performance data useful for debugging
  - + historical data are stored
  - + high scalability and good data visualization
  - + Ganglia Monitoring Agent (gmond)
    - ◗ runs on each node and gathers data every 20 seconds
    - ◗ some parameters useful for alerting too
    - ◗ helps reducing the active Icinga 2 checks

In addition Ganglia and Icinga 2 have been integrated with other data sources, such as SNMP for system information and IPMI for hardware health.

What

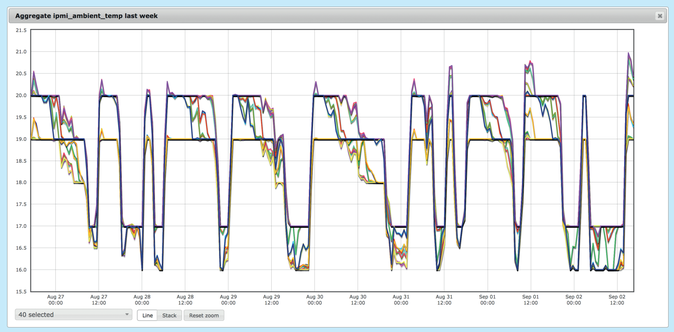
The monitoring system should be able to check up to 100000 health parameters and provide alerts on a selected subset.

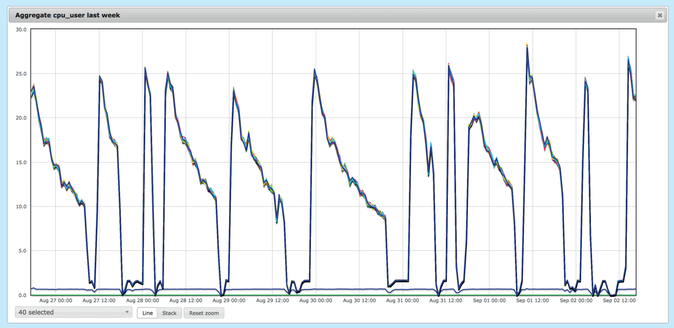
For each node various (~20) hardware and system parameters are monitored:

- ★ hardware:
  - + disk raid status, temperature, fan speed, power suppliers, current and voltages, firmware version...
- ★ system:
  - + cpu, memory and disk usage
  - + configuration aspects as kernel version, pending updates
  - + network: interfaces status, speed and bonding (if present)
  - + services as ntp, http, ssh, mailq

The huge number of nodes and variety of configuration and settings required to set up an automatic way for creating the Icinga 2 configuration files:

- ★ general Icinga 2 templates have been defined
- ★ the in-house Configuration Database (ConfDB) stores the information related to hardware, OS, server types (netbooted or localboot) and specific settings for the Icinga 2 checks, e.g. overriding default thresholds
- ★ SQL queries and regexp, exploiting the node naming convention, are used to extract data and select the right template for the host





This guarantees the full coverage of all the nodes, a per-type configuration and per-host tuning.

A big effort has been put into customising the monitoring of all the significant IPMI sensor information provided by the Baseboard Management Controllers (BMC) of the PCs.

The two Ganglia plots show respectively the ambient temperature and CPU usage of each of the 40 nodes in a rack:

- ➤ the temperature is sent to Ganglia by the IPMI sensors on the nodes
- ➤ the CPU usage is sent to Ganglia by the Ganglia Monitoring Agent running on each node

Monitoring system

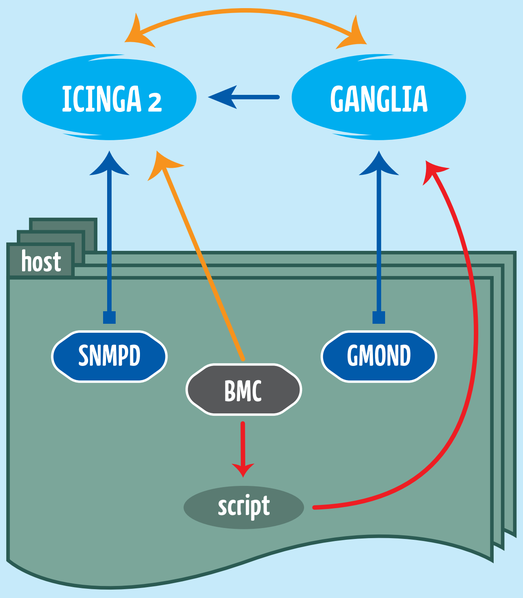
DATA FLOW:

SNMPD
data polled by Icinga 2 ***(blue lines)***

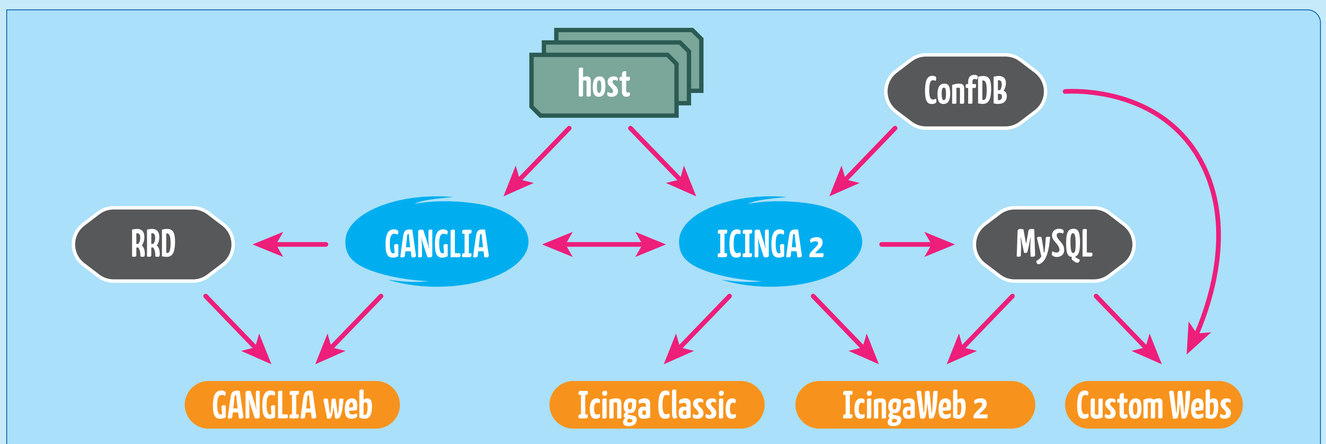
GMOND
data pushed to Ganglia, Icinga 2 can poll Ganglia ***(blue lines)***

BMC
Icinga 2 polls critical data from BMC, stores copy to Ganglia ***(orange lines)***

script polls the host's BMC, pushes complete data to Ganglia ***(red lines)*** (no info if host is down)

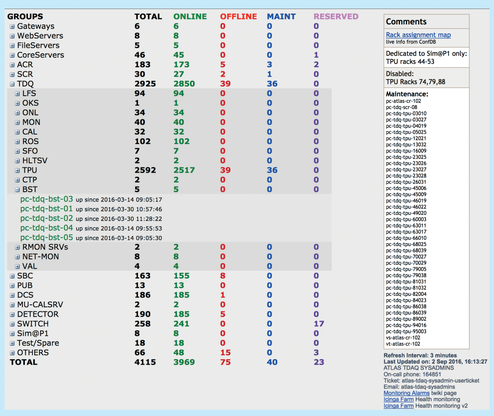


The arrows indicate the data flow



Ganglia and Icinga 2 are the core of the monitoring system: the information from both is used to provide alerts and an historical evolution of the monitored parameters. The data is saved as RRD and in a MySQL database to be used by various web interfaces. The web interfaces show at a glance the status of the farm, from the health status to the resources currently used by data taking.

Farm overview



GROUPS	TOTAL	ONLINE	OFFLINE	MAINT	RESERVED
Gateways	6	6	0	0	0
WebServers	8	8	0	0	0
FileServers	5	5	0	0	0
CoreServers	46	45	0	0	1
ACR	183	173	5	3	2
SCR	30	27	2	1	0
TDQ	2925	2850	39	36	0
LFS	94	94	0	0	0
OKS	1	1	0	0	0
ONL	34	34	0	0	0
MON	40	40	0	0	0
CAL	32	32	0	0	0
ROS	102	102	0	0	0
SFO	7	7	0	0	0
HLTSV	2	2	0	0	0
TPU	2592	2517	39	36	0
CTP	2	2	0	0	0
BST	5	5	0	0	0
pc-tdq-bst-03 up since 2016-03-14 09:05:17					
pc-tdq-bst-01 up since 2016-03-30 10:57:46					
pc-tdq-bst-02 up since 2016-03-30 11:28:22					
pc-tdq-bst-04 up since 2016-03-14 09:55:53					
pc-tdq-bst-05 up since 2016-03-14 09:05:30					
RMON SRVs	2	2	0	0	0
NET-MON	8	8	0	0	0
VAL	4	4	0	0	0
SBC	163	155	8	0	0
PUB	13	13	0	0	0
DCS	186	185	1	0	0
MU-CALSRV	2	2	0	0	0
DETECTOR	190	185	5	0	0
SWITCH	258	241	0	0	17
Sim@P1	8	8	0	0	0
Test/Spare	18	18	0	0	0
OTHERS	66	48	15	0	3
TOTAL	4115	3969	75	40	23

Comments
Rack assignment map
live info from ConfDB
Dedicated to Sim@P1 only:
TPU racks 44-53
Disabled:
TPU Racks 74,79,88
Maintenance:
pc-atlas-cr-102
pc-tdq-scr-08
pc-tdq-tpu-03010
pc-tdq-tpu-03027
pc-tdq-tpu-04019
pc-tdq-tpu-10201
pc-tdq-tpu-13211
pc-tdq-tpu-13332
pc-tdq-tpu-16009
pc-tdq-tpu-23025
pc-tdq-tpu-23026
pc-tdq-tpu-23027
pc-tdq-tpu-23028
pc-tdq-tpu-23031
pc-tdq-tpu-45008
pc-tdq-tpu-45009
pc-tdq-tpu-46019
pc-tdq-tpu-46022
pc-tdq-tpu-49020
pc-tdq-tpu-60003
pc-tdq-tpu-63011
pc-tdq-tpu-63017
pc-tdq-tpu-66010
pc-tdq-tpu-66011
pc-tdq-tpu-68039
pc-tdq-tpu-70027
pc-tdq-tpu-70029
pc-tdq-tpu-79005
pc-tdq-tpu-79038
pc-tdq-tpu-81031
pc-tdq-tpu-81032
pc-tdq-tpu-82004
pc-tdq-tpu-84017
pc-tdq-tpu-84018
pc-tdq-tpu-86039
pc-tdq-tpu-89002
pc-tdq-tpu-94016
pc-tdq-tpu-95003
vs-atlas-cr-102
vt-atlas-cr-102
Refresh Interval: 3 minutes
Last Updated on: 2 Sep 2016, 16:13:27
ATLAS TDAQ SYSADMINS
On-call phone: 164851
Ticket: atlas-tdaq-sysadmin-userticket
Email: atlas-tdaq-sysadmins
Monitoring Alarms twiki page
Icinga Farm Health monitoring
Icinga Farm Health monitoring v2

A custom web page shows an overview of the status of the whole farm.

A "cross-consistency check" custom web page provides information from Icinga, ConfDB and the ATLAS Run status: it displays a quick overview of the resources not in use by ATLAS because of some ongoing hardware or software intervention.

Running ATLAS Partition info: /www/ALL/prod_dyn/sysadmin/check/cc.ATLAS.run updated 2016-09-07 17:35:09.
OKS info: /www/ALL/prod_dyn/sysadmin/check/cc.ALL.defined updated 2016-09-07 17:35:11.

hostname	Icinga		ConfDB HS			TDAQ Part		Issues
pc-tdq-tpu-03010	down	ack=1 maint=0	20	3	maint,sysadmin	oks/0	out	(set Icinga maint)(ok to intervene)
pc-tdq-tpu-03022	up	ack=0 maint=0	20	11	prod,tdaq	oks/1	out	
pc-tdq-tpu-04019	up	ack=0 maint=0	20	11	prod,tdaq	oks/1	out	

Conclusions

The new monitoring system based on Icinga 2 and Ganglia provides the required information and alert notifications.

Our experience shows that a comprehensive and robust monitoring system allows to prevent a lot of hardware and software issues in advance, before they become critical for the ATLAS data taking and the security of the whole system.

Design: Nadia Neri