Towards the integrated ALICE Online-Offline monitoring subsystem

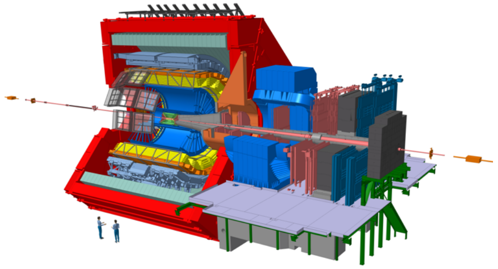
Adam Wegrzynek

for the ALICE Collaboration

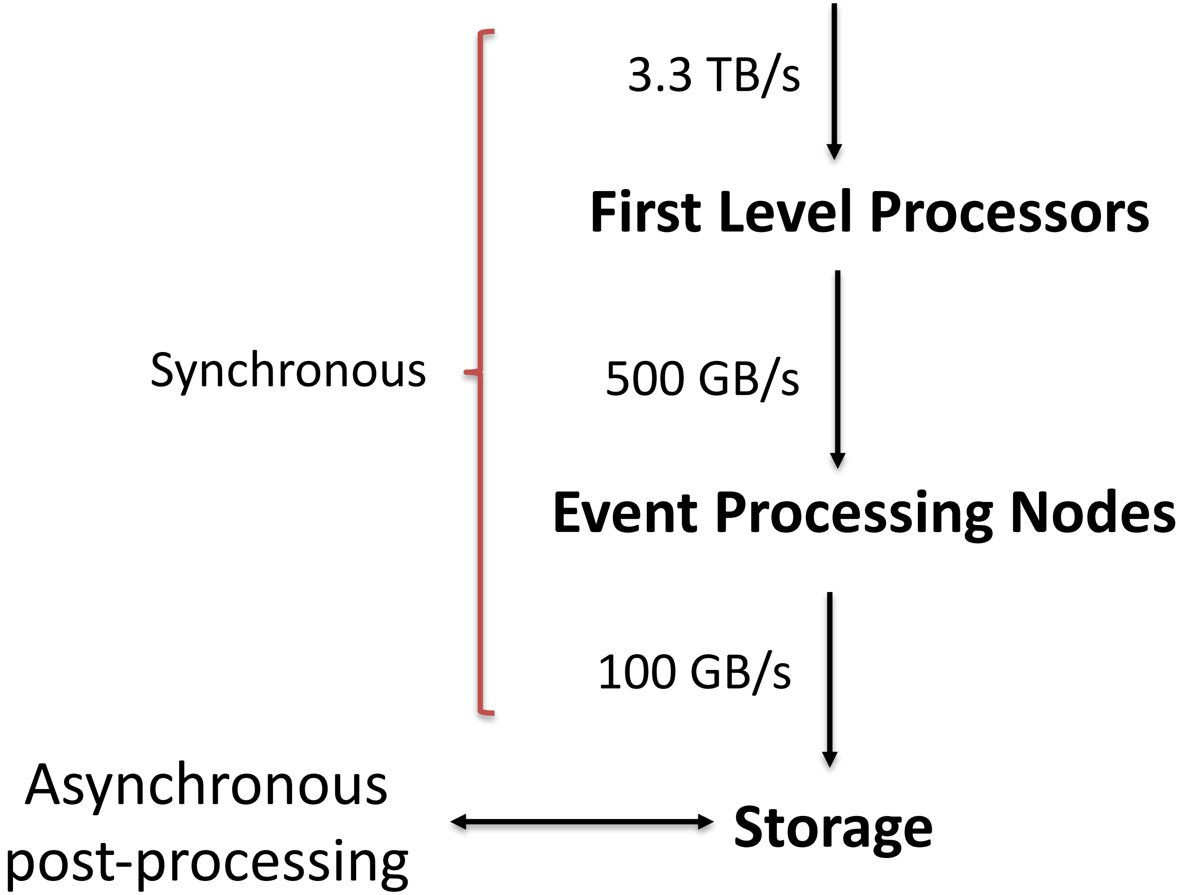
9-13 July 2018 | Sofia, BG

CHEP 2018
Sofia, Bulgaria

ALICE O^2



19 detectors



3.3 TB/s

9000 fibers

First Level Processors

270 nodes

Synchronous

500 GB/s

Event Processing Nodes

1500 nodes

100 GB/s

Asynchronous post-processing

Storage

60 PB storage

Comparison

1.

APACHE Spark InfluxDB

Modular stack

RIEMANN FLUME

2.

MonALISA

MONitoring Agents using a Large Integrated Services Architecture

3.

ZABBIX

- Performance requirements
- Functional architecture
- Experience at CERN

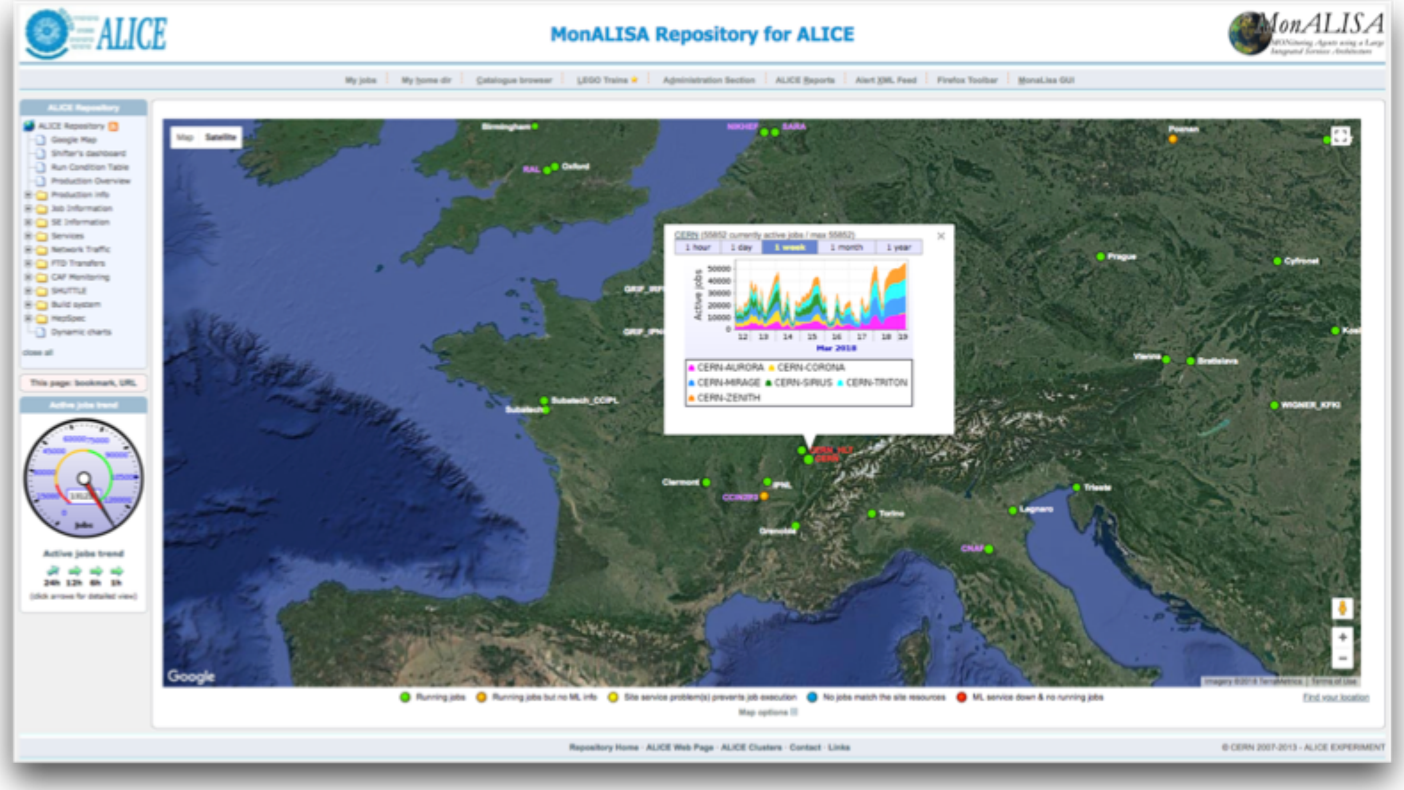
1. *Modular stack*

1. collectd
   - System performance metrics
   - Hardware monitoring
2. FLUME
   - Metric routing
3. APACHE Spark™
   - In memory data processing
4. InfluxDB
   - Time series database
5. Grafana
   - Visualization tool
6. RIEMANN
   - Alarming

- Currently used at INFN Bari, CERN IT

2. MonALISA

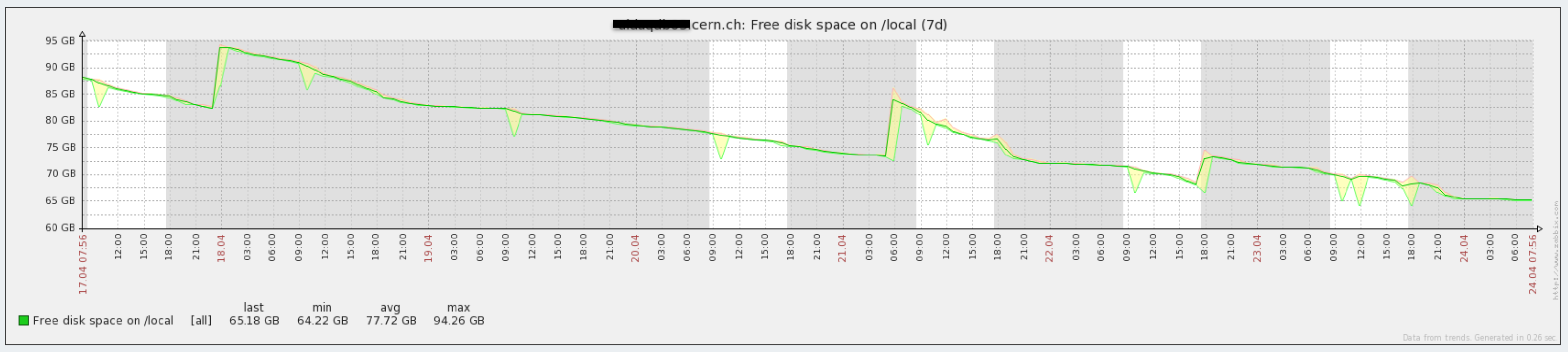
- Distributed data collector infrastructure
- Discovery mechanism
- Aggregation, filtering, alerts
- Real-time data distribution
- In memory buffers
- SQL database
- Currently used by ALICE Offline



Courtesy of **Costin Grigoraş**

3. Zabbix

- Agent-server
- Push via Zabbix protocol
- Community support
- Currently used in ALICE **HLT** and **DAQ** for computing infrastructure monitoring



[redacted]cern.ch: Free disk space on /local (7d)

		last	min	avg	max
🟩 Free disk space on /local	[all]	65.18 GB	64.22 GB	77.72 GB	94.26 GB

Data from trends. Generated in 0.26 sec.

Comparison table (1)

	Modular Stack		MonALISA		Zabbix	
Reference OS (CC7)	🟩	Yes	🟩	Yes	🟩	Yes
Documentation	🟩	Good	🟨	Insufficient	🟩	Good
Support and maintenance	🟩	Yes	🟩	Yes	🟩	Yes
Running in isolation	🟩	Yes	🟩	Yes	🟩	Yes
600 kHz rate	🟩	Yes	🟩	Yes	🟥	No
Scalable >>600 kHz	🟩	Yes	🟩	Yes	🟥	No
Handle 100k sources	🟩	Yes	🟩	Yes	🟥	No
Storage size	🟩	~30 bytes	🟩	~75 bytes	🟨	90-500 bytes

Comparison table (2)

	Modular Stack		MonALISA		Zabbix	
Functional arch.						
System sensors	🟩	Yes	🟩	Yes	🟩	Yes
Metric processing	🟩	Batch and stream	🟩	Stream	🟩	Batch
Historical dashboard	🟩	Yes	🟩	Yes	🟩	Yes
Real-time dashboard	🟧	No (RFC)	🟨	Yes (obsolete)	🟥	No
Alarming	🟩	Yes	🟩	Yes	🟩	Yes
Storage downsampling	🟩	Yes	🟩	Yes	🟩	Yes

Selection

1.

APACHE Spark

InfluxDB

Modular stack

RIEMANN

FLUME

2.

MonALISA

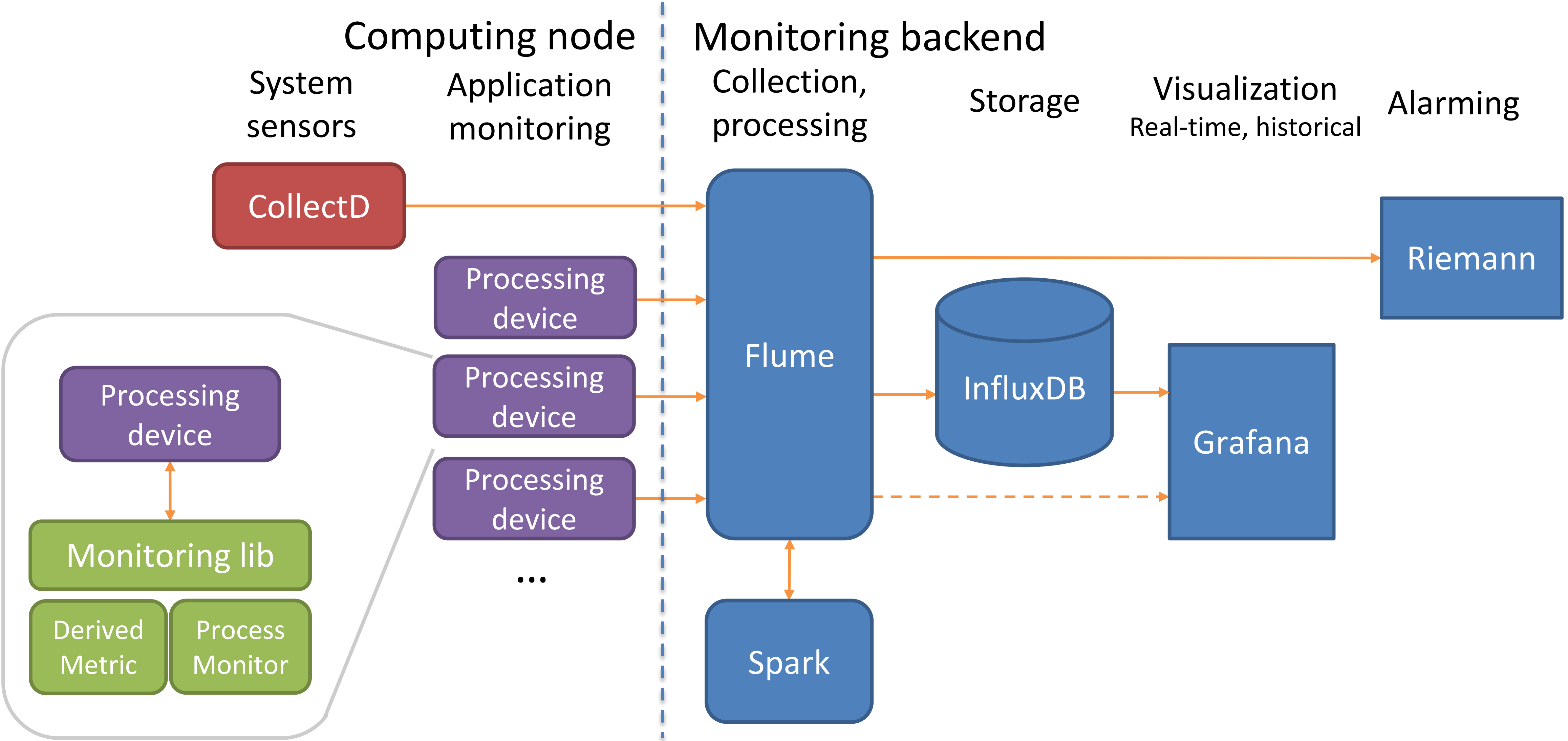
MONitoring Agents using a Large Integrated Services Architecture

Remains for Grid job monitoring

3.

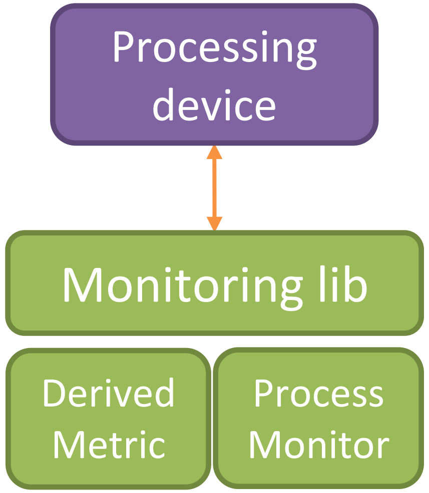
ZABBIX

Modular stack metric flow



Monitoring library

- Push metrics to a backend
- Monitors the process
- Derived metrics
- Tags
- AliceO2Group/Monitoring

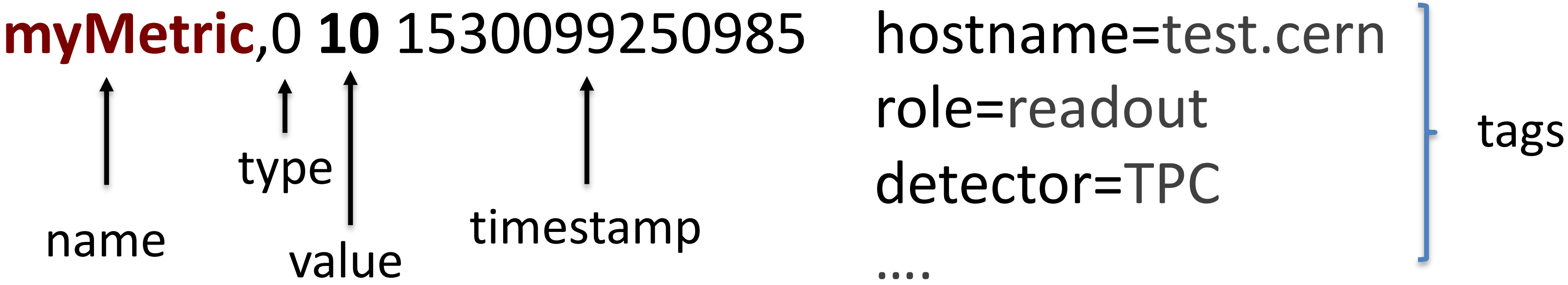


Processing device

Monitoring lib

Derived Metric

Process Monitor



myMetric,0 **10** 1530099250985 hostname=test.cern role=readout detector=TPC ....

name

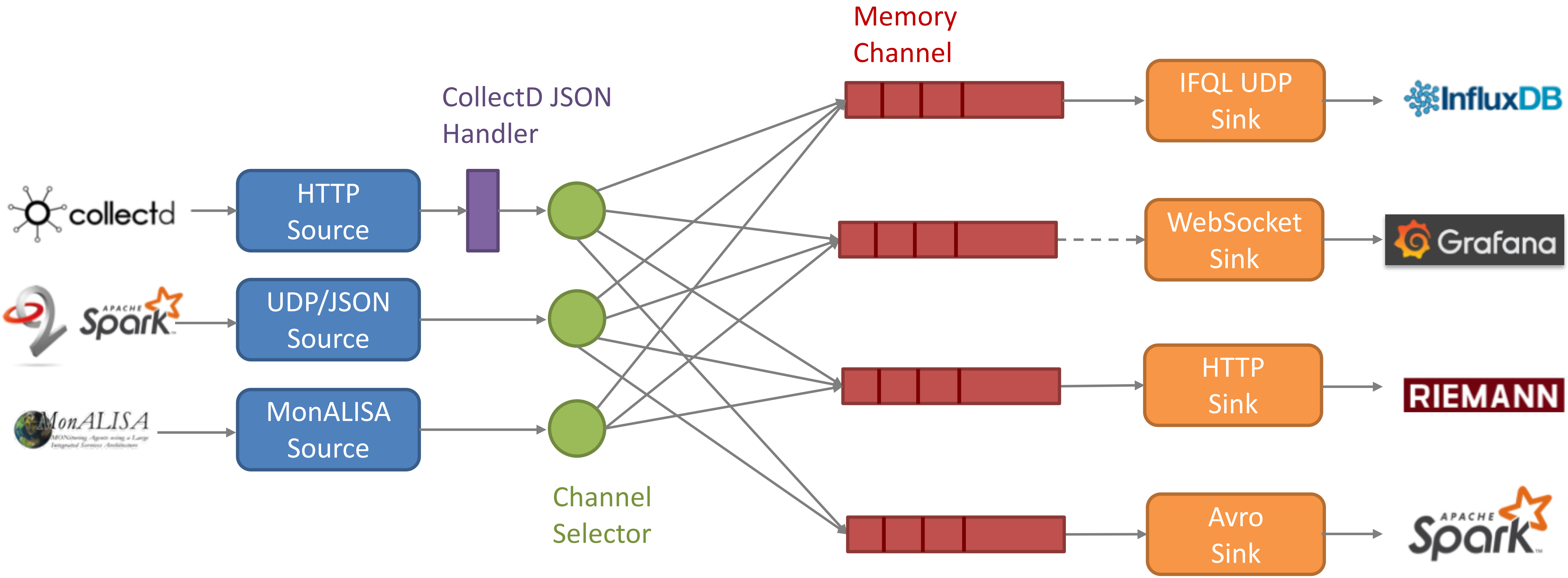
type

value

timestamp

tags

Flume routing



Courtesy of **Gioacchino Vino**

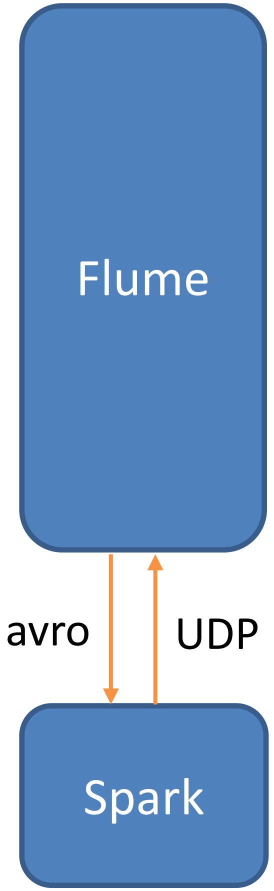
Spark jobs

- Higher level metrics
- Written in Scala
- Operates on Flume events
- Configurable

Readout rate of a detector

↑

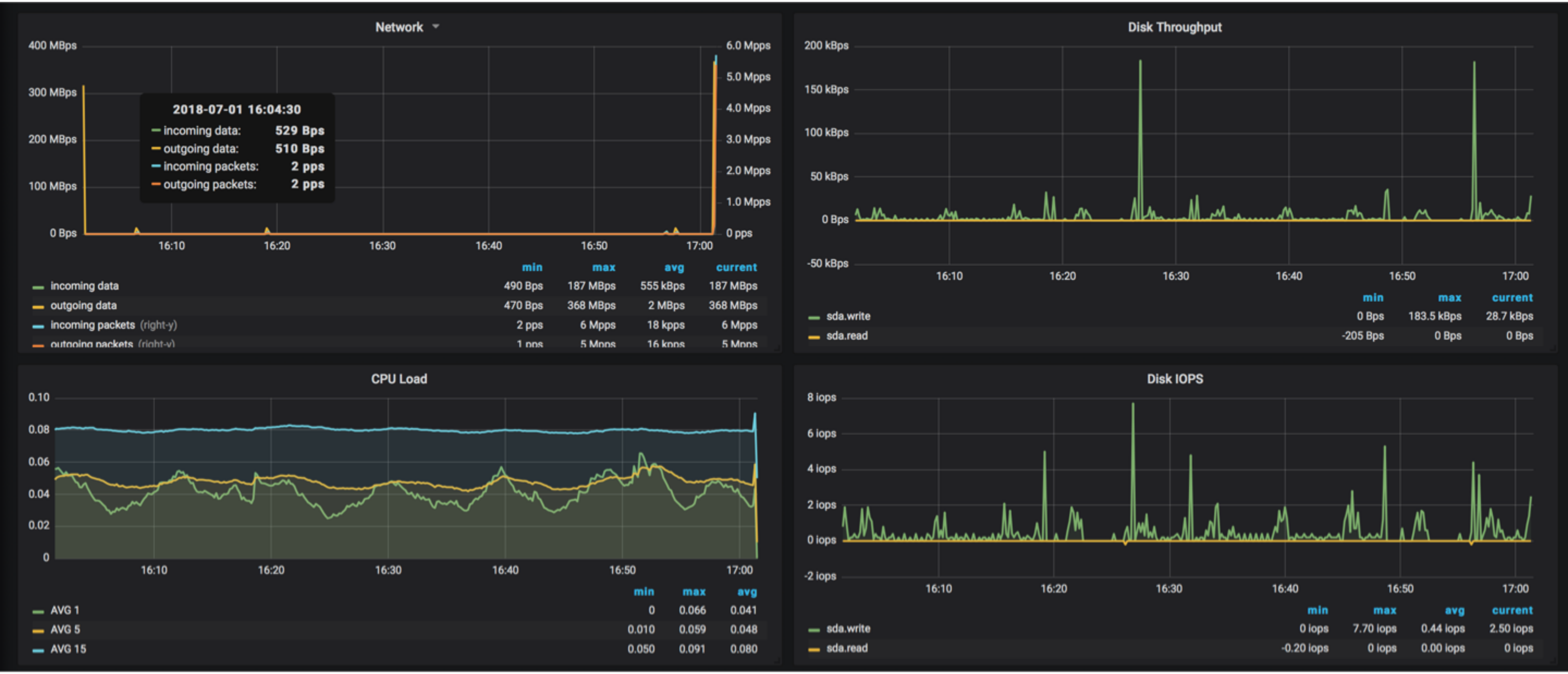
Sum rate of each detector link



InfluxDB timeseries storage

- Up to 700 kHz writes
  - 8 data streams
  - 2x SSD drives RAID0, 25 GbE
- Continuous Queries
  - Downsample high resolution data
  - (merge 12 points into 1 by applying average)
- Retention Policies
  - Drop high resolution data after 30 days
  - Keep low resolution data for 1 year

Grafana



2018-07-01 16:04:30
- incoming data: 529 Bps
- outgoing data: 510 Bps
- incoming packets: 2 pps
- outgoing packets: 2 pps

	min	max	avg	current
incoming data	490 Bps	187 MBps	555 kBps	187 MBps
outgoing data	470 Bps	368 MBps	2 MBps	368 MBps
incoming packets (right-y)	2 pps	6 Mpps	18 kpps	6 Mpps
outgoing packets (right-y)	1 pps	5 Mpps	16 kpps	5 Mpps

	min	max	current
sda.write	0 Bps	183.5 kBps	28.7 kBps
sda.read	-205 Bps	0 Bps	0 Bps

	min	max	avg
AVG 1	0	0.066	0.041
AVG 5	0.010	0.059	0.048
AVG 15	0.050	0.091	0.080

	min	max	avg	current
sda.write	0 iops	7.70 iops	0.44 iops	2.50 iops
sda.read	-0.20 iops	0 iops	0.00 iops	0 iops

Integration with O^2 Software

- Quality Control
- Data Processing Layer

  Evolution of the ALICE Software Framework for LHC Run 3

  Giulio Eulisse, Tuesday 10 July 14:15, Hall 3

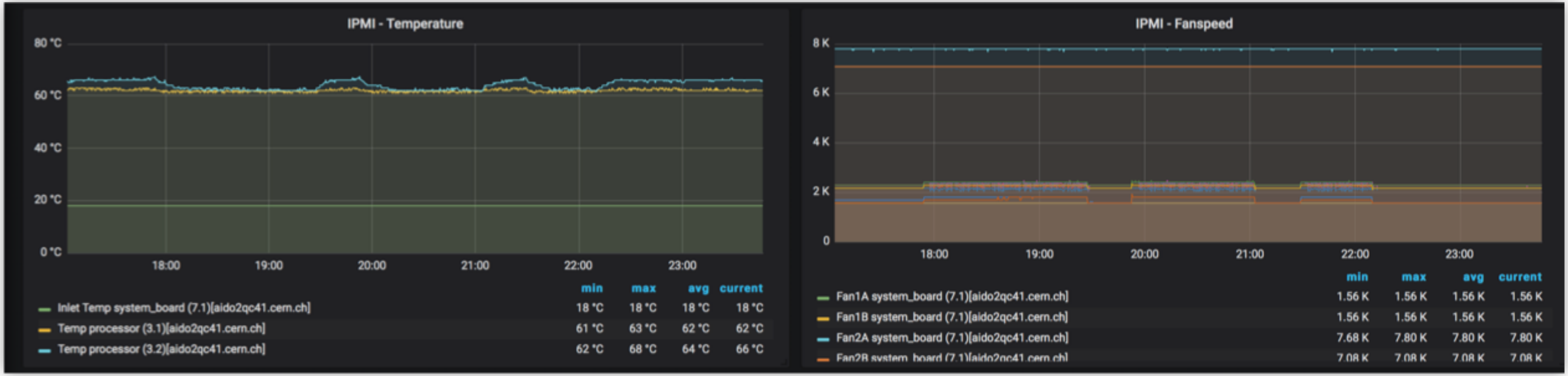
- Readout

  Readout software for the ALICE integrated Online-Offline (O2) system

  Filippo Costa, Thursday 12 July 11:00, Hall 3.1

Conclusion

- 3 options compared
- Modular Stack selected for O^2 farm monitoring
- Defined interfaces between tools
- Deployed in the detector commissioning facilities



IPMI - Temperature

	min	max	avg	current
🟩 Inlet Temp system_board (7.1)[aido2qc41.cern.ch]	18 °C	18 °C	18 °C	18 °C
🟨 Temp processor (3.1)[aido2qc41.cern.ch]	61 °C	63 °C	62 °C	62 °C
🟦 Temp processor (3.2)[aido2qc41.cern.ch]	62 °C	68 °C	64 °C	66 °C

IPMI - Fanspeed

	min	max	avg	current
🟩 Fan1A system_board (7.1)[aido2qc41.cern.ch]	1.56 K	1.56 K	1.56 K	1.56 K
🟨 Fan1B system_board (7.1)[aido2qc41.cern.ch]	1.56 K	1.56 K	1.56 K	1.56 K
🟦 Fan2A system_board (7.1)[aido2qc41.cern.ch]	7.68 K	7.80 K	7.80 K	7.80 K
🟧 Fan2B system_board (7.1)[aido2qc41.cern.ch]	7.08 K	7.08 K	7.08 K	7.08 K

Future steps

- Alarming
  - Define thresholds and patterns
- Grafana real-time data source
  - Display critical metrics in real time
- Sensors to custom hardware
  - Monitor status of custom FPGA board