

SynapSense Wireless Environmental Monitoring System of the RHIC & ATLAS Computing Facility at BNL

K. Casella¹, E. Garcia¹, R. Hogue¹, C. Hollowell¹, W. Strecker-Kellogg¹,
A. Wong¹, A. Zaytsev¹

¹ Brookhaven National Laboratory, Upton, NY 11973, USA

E-mail: alezayt@bnl.gov, kac@bnl.gov

Abstract. RHIC & ATLAS Computing Facility (RACF) at BNL is a 15000 sq. ft. facility hosting the IT equipment of the BNL ATLAS WLCG Tier-1 site, offline farms for the STAR and PHENIX experiments operating at the Relativistic Heavy Ion Collider (RHIC), the BNL Cloud installation, various Open Science Grid (OSG) resources, and many other small physics research oriented IT installations. The facility originated in 1990 and grew steadily up to the present configuration with 4 physically isolated IT areas with the maximum rack capacity of about 1000 racks and the total peak power consumption of 1.5 MW. In June 2012 a project was initiated with the primary goal to replace several environmental monitoring systems deployed earlier within RACF with a single commercial hardware and software solution by SynapSense Corporation based on wireless sensor groups and proprietary SynapSense™ MapSense™ software that offers a unified solution for monitoring the temperature and humidity within the rack/CRAC units as well as pressure distribution underneath the raised floor across the entire facility. The deployment was completed successfully in 2013. The new system also supports a set of additional features such as capacity planning based on measurements of total heat load, power consumption monitoring and control, CRAC unit power consumption optimization based on feedback from the temperature measurements and overall power usage efficiency estimations that are not currently implemented within RACF but may be deployed in the future.

1. Introduction

RHIC & ATLAS Computing Facility (RACF) [1] at Brookhaven National Laboratory (BNL) [2] is a 15000 sq. ft. facility hosting the IT equipment of the BNL ATLAS Tier-1 site [3] within the Worldwide LHC Computing Grid (WLCG) [4] infrastructure, offline farms for the STAR [5, 6] and PHENIX [7, 8] experiments operating at the Relativistic Heavy Ion Collider (RHIC) [9, 10], the BNL Cloud installation [11], various Open Science Grid (OSG) [12] resources, and many other small physics research oriented IT installations. The facility originated in 1990 and grew steadily up to the present configuration with four main physically isolated IT areas with the maximum rack capacity of about 1000 racks plus several auxiliary areas. Each of these isolated areas contains equipment grouped by functionality, purpose and ownership by the particular project and/or collaboration.

Currently the main IT areas are:

1. BCF (originally the Brookhaven Computing Facility), the oldest area in the data center.
2. RCF (RHIC Computing Facility) serving computing and storage needs of STAR and PHENIX experiments at RHIC collider.
3. Sigma-7 area primarily devoted to storage systems.



4. CDCE (Computing Data Center Expansion) devoted to the ATLAS WLCG Tier-1 site, the newest area in the data center.
5. Auxiliary areas such as NetLabs, parallel Supercomputer areas [13], etc...

In its present configuration the RACF facility has the following integral characteristics:

- More than 400 racks deployed across all the areas.
- 1.3 MW diesel generator plus 2 flywheel UPS systems; 1 MW battery UPS.
- 1.5 MW of peak total power consumption.
- 34k HT CPU cores plus 72 TB RAM in Linux farms.
- 1 PB in centralized NFS storage based on BlueArc/HDS storage solutions [14].
- 23 PB of distributed storage based on dCache [15] and XRootD [16].
- 66k slots in robotic tape silos: 35 PB in HPSS [17] storage system.

All the areas of the data center are making use of the room level air cooling systems provided with humidity control supplied by one or several of the following cooling subsystems:

- Liebert downflow CRAC units [18] distributed among the area of every IT room (central BNL-wide chilled water supply is used),
- Central building-level AC units in the basement (nowadays considered legacy equipment),
- Liebert racktop cooling units [19] based on pumped refrigerant technology are used in certain high density areas.

2. The Need for an Environmental Monitoring System

The RACF facility requires a robust environmental monitoring system covering the area of all the IT rooms for the following two primary reasons.

First, our facility mostly relies on a pressurized raised floor based solution for distributing cooled air in all the areas, thus maintaining a uniform pressure differential on the entire area of the facility is important for ensuring the high efficiency of the room-level cooling systems. Taking into consideration the complexity of the configuration of the raised floor plates equipped with the airflow grates, and the fact that the configuration changes frequently as IT equipment gets added/removed in various areas of the data center, one can see that it's difficult to maintain the uniform pressure differential distribution without a system that measures it on a regular basis.

The second reason is due to the fact that RACF is equipped with a redundant power supply and cooling infrastructure, and these redundancy systems are operating independently of one another. Thus a temperature monitoring system with a rapid response rate is required in order to avoid the risk of damaging IT equipment in case of infrastructure failures. Such risks could be divided into two broad categories.

First of all, there are chances of local and/or room level overheating in certain IT areas with relatively small air volume and/or large portions of IT equipment on UPS. In case of power failure the IT load switches to UPS while the cooling system may go down, causing fast overheating that can result in a permanent damage of the IT equipment deployed in that area. Such incidents are usually prevented by implementing an automated shutdown mechanism that uses the temperature measurements provided by an environmental monitoring system, which cuts power to the IT equipment located in the overheating racks. Since in case of a cooling system failure in the high density areas the ambient temperature can increase at a rate of 1 °C/min, fast response time of the monitoring system and the automated shutdown mechanisms is needed.

The second category includes severe local overheating (above 60-70 °C ambient temperature inside a rack) of the racks containing storage equipment caused by a racktop cooling unit failure or temporary raised floor pressure differential misbalance (for instance, caused by removal of a section of the raised floor in the vicinity), that may trigger imminent disk failures in the RAID systems (especially for the systems loaded with HDDs that have been in production for 2-3 years already). That may cause a significant reduction in the storage system performance for the period of RAID group recovery and even an unrecoverable RAID system degradation. Thus, observing only the

ambient temperature in an IT room is not sufficient for preventing such issues, and the environmental monitoring system must have a high spatial resolution allowing one to detect temperature anomalies (ideally) in every rack in the facility.

The following additional requirements must be satisfied by the environmental monitoring system best suited for RACF data center:

- Ability to analyze historical data in order to perform retrospective analysis of maintenance and disaster recovery operations and their impact on environmental conditions in the data center.
- Ease of reconfiguration and reinstallation of the sensors matching the dynamic nature of the RACF environment where rack movements and row reconfigurations do happen frequently.

3. Historical Perspective and Considerations for the New Environmental Monitoring System

The following systems were deployed and used within RACF computing facilities over the last several years until 2011:

- Ambient temperature and humidity sensors with local displays deployed in every room (no central data gathering/alarm system yet).
- Embedded CRAC unit sensors provided with local alarms (no central information gathering implemented so far).
- DigiTemp solution [20] based ambient rack temperature sensors: provided with central data gathering mechanism, but unable to supply the data with high spatial resolution.
- IPMI [21] based temperature information gathering: provides high spatial granularity but has poor absolute precision and reduced sensitivity to the room-level environmental changes. IPMI data is sufficient for implementing host level automatic emergency shutdown mechanisms, but those are normally used only as a last resort.

The following architectural possibilities were considered for the new unified and centralized RACF environmental monitoring system. First of all, the solution that would seem the easiest was to simply extend the existing sensor groups provided with the Ethernet connectivity (factor of 10 increase of sensor count is needed), but on the scale of RACF that would require deploying too much of the dedicated cable infrastructure. The second option was to deploy new groups of sensors provided with wired loop topology interlinks and central controllers (such as APC/NetBotz monitoring systems [22]), but they again would require a dedicated cable infrastructure intersecting multiple racks which would significantly limit the flexibility of rack movements in the facility.

The third option also available on the market was to make use of fiber optic based monitoring/data gathering solutions such as Raman [23, 24] and Brillouin [25, 26] distributed fiber based thermometers. While demonstrating a remarkably high spatial resolution, such systems require a dedicated fiber optics infrastructure intersecting multiple racks plus precision placement of the fibers within the racks. Another drawback is due to the limitation on the number of junctions that are allowed on a fiber line due to signal attenuation restrictions. Moreover, the systems with high spatial resolution (better than 30 cm) are rather expensive. Last but not least, the forth option we considered was to deploy a wireless network based solution such as one provided by SynapSense Corporation [27].

Upon careful evaluation of these options the decision was made in mid-2012 to choose the wireless solution by SynapSense Corporation as the most price/performance-efficient solution for our environment for deployment in all the areas of RACF facility. It is also noteworthy that a similar system [28] was successfully deployed in Lawrence Berkeley National Laboratory (LBNL) [29] several years ago, thus the solution chosen by BNL was considered tested within an organizational and operational environment similar to ours.

4. SynapSense™ Monitoring System Architecture

The SynapSense™ wireless environmental monitoring system consists of both hardware and software components. From the hardware perspective it consists of the following main building blocks:

1. SynapSense™ Gateway relays data between wired and wireless networks. This wireless transceiver communicates with environmental sensors to send data to the SynapSoft

management software. Depending on the amount of material between the Gateway unit and the sensor base stations obstructing the signal the range of stable wireless communication can reach up to 50 meters.

2. SynapSense™ ThermaNode measures temperature and humidity from the data center racks, CRACs/CRAHs, plenums, and other environmental areas. Battery operated and wireless; each provided with the embedded humidity sensor.
3. SynapSense™ Pressure Node measures air pressure differences between two points in the data center (subfloor and overhead plenums). The unit is battery operated and wireless.

SynapSense™ ThermaNode can be provided with various configurations of the temperature sensor assemblies (the sensors assembly are attached to the ThermaNode with analogue wired lines), of which the following three are of primary interest to us:

1. Front/rear doors plus subfloor assembly for racks (7 temperature sensors),
2. So called SynapSense™ LiveImaging™ temperature sensor assembly (3 sensors),
3. CRAC unit sensor assembly (2 temperature sensors).

The maximum expected lifetime of batteries deployed in ThermaNodes and Pressure Node is reaching up to 10 years according to the vendor's specifications. Since these units are making use of the standard AA form-factor batteries, they can be changed easily when needed.

Software-wise there are the following main components in the system. First is the server-side database based on commercial MySQL [30] and data gathering SynapSense™ SynapSoft™ software designed to be deployed on top of a Microsoft Windows Server based environment. The display and management of sensor data is natively done via the SynapSense™ Web Console™ application. Furthermore, the capability to generate and view near-real-time 2D maps of the sensor-monitored environment through use of a thermal map style color gradient overlaid onto the data center layout image is provided by the SynapSense™ LiveImaging™ product. Finally, SynapSense™ MapSense™ is a layout and configuration tool used to map the physical deployment of sensors installed on data center equipment to an equivalent software model displayed in Web Console™. Please refer to [31-33] for more information on various components of the system.

5. Deployment and Production Operations of the SynapSense™ Monitoring System in RACF

The deployment of the new wireless monitoring system was carried out in 3 stages. Stage 1 (June 2012): the central SynapSense™ server infrastructure was deployed along with the first SynapSense sensor group in the RCF area. Stage 2 (December 2012): the second SynapSense™ sensor group was deployed in the Sigma-7 area. Stage 3 (August 2013): completion of the monitoring system deployment: BCF and CDCE area are covered plus a major SynapSense™ software upgrade was performed. Most of the installation operations were performed by a group of SynapSense engineers, save the deployment in Sigma-7 area that was handled exclusively by the RACF staff.

In the final configuration the system consists of six wireless gateways, 152 ThermaNode units, and 926 temperature, humidity, and pressure differential sensors (so called "sense points") reporting their measurements to the central database every five minutes. Sensor readings are being pushed to the central database in an asynchronous fashion with a jitter of about 1 minute, so the sensor reports are being gathered within time windows of 45-90 seconds that occur every five minutes. This way the spike load on the database server is reduced substantially compared to an alternative scenario of getting all the sensor readings simultaneously.

The current layout of the system is shown in Fig. 1. Starting from third quarter of 2012 the system was complemented by the addition of a Scientific Linux 6.x x86_64 based server provided with an instance of an open source MySQL database and a set of scripts performing an offload of the new sensor readings into this secondary database every minute in a format that can be directly used by the automated IT load shutdown mechanisms that were previously developed within RACF (e.g. for the RCF area). The frequency of the data offloading attempts is intentionally set higher than the characteristic data gathering frequency of the sensors in order to better deal with the asynchronous nature of the data gathering mechanism used by SynapSense™ central services.

Two examples of the 2D temperature and pressure differential maps of the facility generated by the SynapSense™ LiveImaging™ software are shown in Fig. 2 and Fig. 3 respectively.

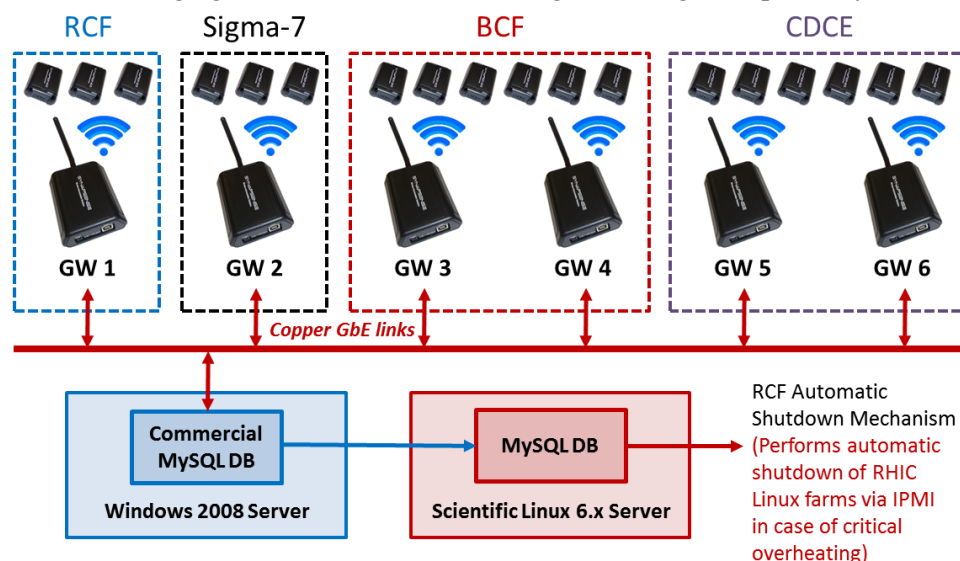


Figure 1. Overall layout of the SynapSense™ wireless data gathering system, data archival and extraction mechanism and the automatic emergency shutdown mechanism that is in place for certain areas of the RACF facility.

6. Summary and Future Developments

Full scale deployment of the SynapSense™ environmental monitoring system was performed in the RACF data center during the period of June 2012 – August 2013. In the present configuration the system has more than 150 base stations provided with more than 920 temperature, humidity, and pressure sensors reporting to the central servers every five minutes resulting in 0.27M readings per day stored in the centralized database. The integral cost of the system is not exceeding the cost of two racks of equipment typical for RACF Linux farms.

Experience gathered with the system so far indicates that we achieved the necessary level of time and spatial resolution that allows us to:

- Monitor the rack-by-rack cooling issues such as racktop cooling unit failures.
- Detect IT room level issues such as raised floor pressure differential imbalance, AC unit failures, and obstructed airflow issues.
- Observe and perform retrospective analysis of evolution of environmental conditions within the entire RACF data center.

SynapSense monitoring system is considered reliable enough to supply the input data for the mechanism that performs automatic shutdown of the IT load in certain areas of RACF data center in case of localized and/or room level overheating. The mechanism is used in production since mid-2012 for the RCF area. There is also a potential of extending the SynapSense™ monitoring system to include power consumption monitoring for all the CRAC and PDU devices in the facility and utilizing the SynapSense™ ActiveControl™ features [34], thus providing near-real-time estimates for the power usage effectiveness (PUE) factor of the RACF data center and the necessary means to optimize it.

References

- [1] O. Rind, A. Wong *et. al.*, BNL RHIC/ATLAS Computing Facility Site Report. HEPiX Fall 2013 Workshop, October 28 – November 1, 2013, University Of Michigan, Detroit, USA
<https://indico.cern.ch/getFile.py/access?contribId=16&sessionId=0&resId=0&materialId=slides&confId=247864>
- [2] Official website of BNL: <http://www.bnl.gov>

- [3] BNL ATLAS Tier-1 WLCG Site: <https://www.racf.bnl.gov/experiments/usatlas/>
- [4] Worldwide LHC Computing Grid: <http://cern.ch/wlcg/>
- [5] STAR Detector Experiment at RHIC: <http://www.star.bnl.gov>
- [6] J. Adams *et al.*, Experimental and Theoretical Challenges in the Search for the Quark Gluon Plasma: The STAR Collaboration's Critical Assessment of the Evidence from RHIC Collisions. Nucl. Phys. A757, 102 (2005)
- [7] PHENIX Detector Experiment at RHIC: <http://www.phenix.bnl.gov>
- [8] K. Adcox *et al.*, PHENIX detector overview. NIM A499 469-479 (2003)
- [9] Relativistic Heavy Ion Collider (RHIC): <http://www.bnl.gov/rhic/>
- [10] H. Hahn *et al.*, The RHIC design overview. NIM A499 245-263 (2003)
- [11] J. Caballero, J. Hover *et al.*, Cloud activities at BNL. US ATLAS T2/3 Workshop, March 11-14, 2013, Indianapolis, Indiana, USA: <https://indico.fnal.gov/getFile.py/access?contribId=51&sessionId=18&resId=0&materialId=slides&confId=5610>
- [12] Open Science Grid (OSG): <http://www.opensciencegrid.org>
- [13] New York Blue/L and P, BNL Blue Gene/Q Supercomputers: <http://www.bnl.gov/bluegene/>
- [14] Hitachi Unified Storage and Hitachi NAS Platform:
<http://www.hds.com/products/file-and-content/network-attached-storage/>
- [15] dCache Storage System: <http://www.dcache.org>
- [16] XRootD Project: <http://xrootd.slac.stanford.edu>
- [17] High Performance Storage Systems (HPSS): <http://www.hpss-collaboration.org>
- [18] Liebert CW Chilled Water-based AC units: <http://www.emersonnetworkpower.com/en-US/Products/PrecisionCooling/LargeRoomCooling/Pages/LiebertCWChilledWater-basedPrecisionCooling26-181kW.aspx>
- [19] Liebert Refrigerant-based Racktop Cooling Modules: <http://www.emersonnetworkpower.com/en-EMEA/Products/PrecisionCooling/HighDensityModularCooling/Refrigerant-Based/Pages/Liebert-XDV.aspx>
- [20] DigiTemp Sensor Systems for Linux: <http://www.digitemp.com>
- [21] Intelligent Platform Management Interface (IPMI):
<http://www.intel.com/content/www/us/en/servers/ipmi/ipmi-home.html>
- [22] APC/NetBotz Security and Environmental Monitoring Systems:
<http://www.apc.com/products/category.cfm?id=17>
- [23] Optical Fibre based Distributed Sensor for Temperature Measurement:
<http://www.igcar.ernet.in/benchmark/Tech/19-tech.pdf>
- [24] C. Pandian *et al.*, Single-fiber grid for improved spatial resolution in distributed fiber optic sensor. Optics Letters, Vol. 35, Issue 10, pp. 1677-1679 (2010)
- [25] F. Rodríguez-Barrios *et al.*, Distributed Brillouin Fiber Sensor Assisted by First-Order Raman Amplification:
http://infoscience.epfl.ch/record/150155/files/JLT_28_2162_2010_1st%20order%20Raman.pdf
- [26] L. Zou *et al.*, Distributed fiber Brillouin strain and temperature sensor with centimeter spatial resolution by coherent probe-pump technique:
http://net04.isis.umanitoba.ca/activeshm/ReferencePage/Reference/Bao/spie5855_zou_a.pdf
- [27] SynapSense Corporation: <http://www.synapsense.com>
- [28] Dale Sartor *et al.*, Saving Energy in LBNL's Data Centers:
http://fimsinfo.doe.gov/Workshop/2010/Sartor_Data_Center.pdf
- [29] Official website of LBNL: <http://www.lbl.gov>
- [30] MySQL Enterprise Edition: <http://www.mysql.com/products/enterprise/>
- [31] http://www.etcc-ca.com/images/stories/docs/Open_Forum/8_SynapSense_Presentation_for_SMUD.pdf
- [32] SynapSense™ Products/Hardware: http://www.synapsense.com/?page_id=38
- [33] SynapSense™ Products/Software: http://www.synapsense.com/?page_id=40
- [34] SynapSense™ ActiveControl™ Solutions: http://www.synapsense.com/?page_id=27

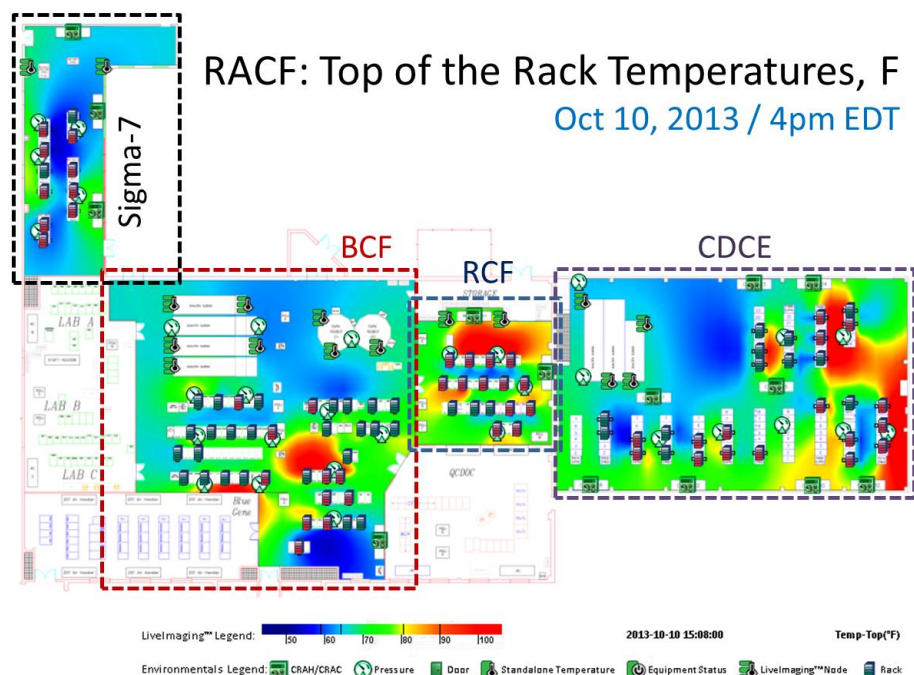


Figure 2. Snapshot of the 2D heat map of the RACF facility for the designated moment in time generated by SynapSense™ LiveImaging™ software. Top of the rack sensor data is used here. Four main IT areas of the facility are highlighted (not a part of the original image).

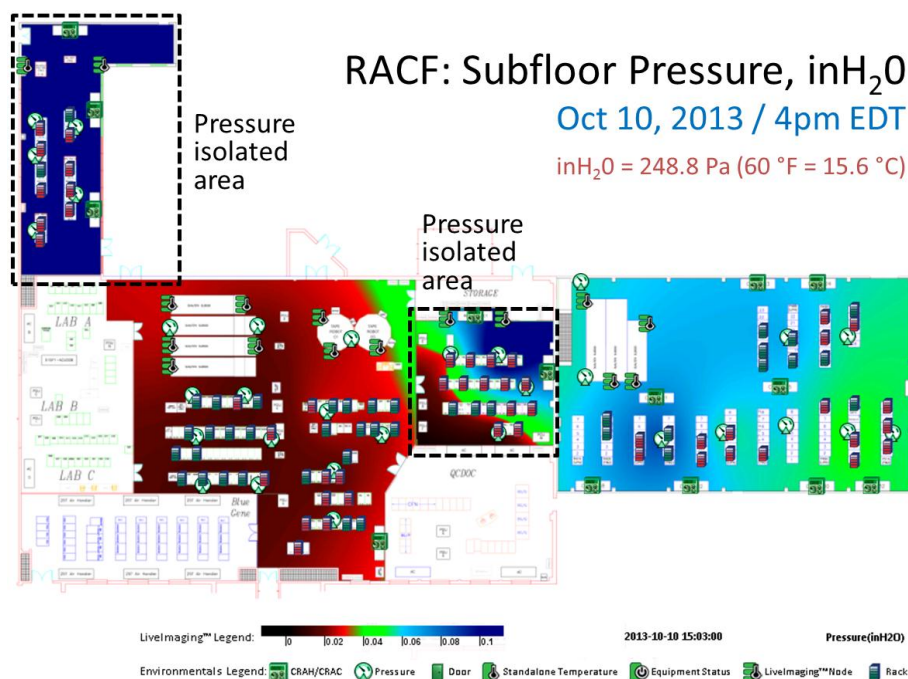


Figure 3. Snapshot of the 2D raised floor pressure differential map of the RACF facility for the designated moment in time generated by SynapSense™ LiveImaging™ software. Pressure isolated areas are shown here as configured in the map of the facility (not a part of the original image).