INL/EXT-11-23408
Rev. 2

INL Control System Situational Awareness Technology Final Report 2013

Gordon Rueff
Bryce Wheeler
Todd Vollmer
Tim McJunkin

January 2013

The INL is a U.S. Department of Energy National Laboratory
operated by Battelle Energy Alliance

PROJECT TEAM: INL NSTB PROGRAM TEAM

INL Management:
DOE-ID Project Lead: John Yankeelov
INL Management Liaison: Curtis St Michel
Program Manager: David Kuipers

Technical Team:
Program Cyber Researcher Support Team:
Jared Verba, Corey Thuen, Ken Rohde, Robert Erbes, Kent Kvarfordt, James Thomas, Larry Wellman, Ann Egger

Program Researcher Support Team:
John Buttles, Eric Larsen, Mark McKay, Karen Miller

Academic Team:
Milos Manic (Univ of Idaho), Grad Students (Univ of Idaho)

Program Support:
Program Financial Consultant: Ben Watts
Program Legal Consultant: Rick Evans
Senior Writer: Zack Adams
Administrative Support: Karen Daniel, Julie Irving
Web Support: Shad Staples, Desiree Reagan, David Loynd
Commercialization: Mark Kaczor, Charity Follet, Kathleen Bohachek

INL/EXT-11-23408
Rev. 2

INL Control System INL Control System Situational Awareness Technology Final Report 2013

Gordon Rueff
Bryce Wheeler
Todd Vollmer
Tim McJunkin

January 2013

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Office of Electricity Delivery and Energy Reliability
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

ABSTRACT

The Situational Awareness project is a comprehensive undertaking of Idaho National Laboratory (INL) in an effort to produce technologies capable of defending the country's energy sector infrastructure from cyber attack. INL has addressed this challenge through research and development of an interoperable suite of tools that safeguard critical energy sector infrastructure. The technologies in this project include the Sophia Tool, Mesh Mapper (MM) Tool, Intelligent Cyber Sensor (ICS) Tool, and Data Fusion Tool (DFT). Each is designed to function effectively on its own, or they can be integrated in a variety of customized configurations based on the end user's risk profile and security needs.

EXECUTIVE SUMMARY

Securing the country's energy sector infrastructure from cyber-attack is critical to the well-being of the American people and is a central focus to the Department of Energy's (DOE) Office of Electricity Delivery and Energy Reliability (OE) Cybersecurity for Energy Delivery Systems (CEDS) program. The DOE program aims to enhance the reliability and resilience of the nation's energy infrastructure by reducing the risk of energy disruptions due to cyber-attacks. As part of the Situational Awareness project, INL has conducted research and developed an interoperable suite of tools to help industry defend America's critical infrastructure from cyber attack.

The **Sophia Tool** provides users a thorough view of their control system and wired sensor networks, allowing a detailed review of conversations that are occurring. During the first quarter of FY 2013, Sophia proceeded by expanding the beta test and working toward meeting internal requirements for Sophia to be licensed to a commercialization partner. Currently Sophia software is going out to interested parties for evaluation.

The **Mesh Mapper (MM) Tool** was intended to collect the routes taken by Supervisory Control and Data Acquisition (SCADA) system Wireless Mesh Network (WMN) data messages and track them in such a way the operator can readily observe any abnormal behavior of wireless sensor networks. There was no work done on the Mesh Mapper tool in FY 2013 Q1. A No-Go decision was made by the project team and DOE-OE CEDS PM after the Informal Design Review in FY 2012 determined that current technology did not adequately implement features to support this research.

The **Intelligent Cyber Sensor (ICS) Tool** distinguishes between component failure and cyber security incidents and monitors the overall health of a system. ICS work in the first quarter of FY 2013 focused on reducing false positives and the implementation of commercial friendly features.

The **Data Fusion Tool (DFT)** identifies, reduces, and characterizes data, providing integrated situational awareness of the cyber and operational health of the control and sensor system. During the first quarter of FY 2013, the Data Fusion Tool (DFT) development was continued primarily to enhance integration with the Intelligent Cyber Sensor (ICS).

The FY 2011 annual report comprehensively described each of the above tools or technologies; the FY 2011 and FY 2012 annual reports provided yearly task completion status; and on August 29, 2012, the interoperability of the tools in the Situational Awareness Suite was demonstrated to DOE-OE and industry representatives.

This report provides an update on work completed on the Situational Awareness technologies since the FY 2012 annual report in October and provides final reporting per the requirements of the project work scope, INL Control System Situational Awareness Technology proposal, July 2010, Task 8.0, "Final Report," Page 32.

CONTENTS

FIGURES

TABLES

ACRONYMS

ADMT	Advanced Data Mining Technique
AP-title	Access Point Title
CEDS	Cybersecurity for Energy Delivery Systems
CI	Computational Intelligence
CR	Cell Relay
DFT	Data Fusion Tool
DOE-OE	Department of Energy Office of Electricity Delivery and Energy Reliability
FLA	Fujitsu Laboratories of America, Inc.
GUI	Graphical User Interface
ICIS	Instrumentation, Control and Intelligent Systems
ICS	Intelligent Cyber Sensor
INL	Idaho National Laboratory
IP	Internet Protocol
LDRD	Laboratory Directed Research and Development
MM	Mesh Mapper
NetAPT	Network Access Policy Tool
NSTB	National SCADA Test Bed program
R&D	Research and Development
SCADA	Supervisory Control and Data Acquisition
SPADUC	SCADA Protocol Anomaly Detection
TD	Technical Deployment
WMN	Wireless Mesh Network

INL Control System Situational Awareness Technology Final Report 2013

1. SITUATIONAL AWARENESS LAB CALL PROJECT

The INL Situational Awareness Technology project is funded through the Department of Energy's (DOE) Office of Electricity Delivery and Energy Reliability (OE) Cybersecurity for Energy Delivery Systems (CEDS) Research and Development (R&D) Program. The DOE program aims to enhance the reliability and resilience of the nation's energy infrastructure by reducing the risk of energy disruptions due to cyber-attacks. The project is a two year effort, started in FY 2011, of R&D tasks related to improving industry situational awareness of energy sector control systems cybersecurity through development of new interoperable tools.

2. SOPHIA PROJECT

The Sophia Tool is a software development research effort to create a new tool for fingerprinting and monitoring Supervisory Control and Data Acquisition (SCADA) systems. Sophia provides users with reliable information for decision-making that enhances the security and resilience of their SCADA system. The Sophia concept is designed as a passive, real-time tool for inter-device communication discovery and monitoring of the active elements in a SCADA system.

Complete details on the capabilities and progress of the Sophia tool are provided in INL Control System Situational Awareness Technology Annual Reports for FY 2011 and FY 2012.

2.1 FY 2013 Q1 Sophia Update

During the first quarter of FY 2013, Sophia proceeded by expanding the beta test and working toward meeting internal requirements for Sophia to be licensed to a commercialization partner. The Sophia beta testing period was extended until December 31, 2012, during which time 12 new beta testers joined the program. The final beta software package was generated on October 23, 2012. Internal commercialization requirements included soliciting interested parties, documenting QA procedures to meet DOE requirements, and acquiring new copyrights on the Sophia code base. Currently Sophia software is going out to interested parties for evaluation. A patent for the Sophia concept has been registered.

2.2 Summary

2.2.1 Accomplishments

Sophia added 12 new beta testers in the first quarter of FY 2013, and finished INL procedures for meeting DOE O 414.1 D, Quality Assurance. There were 44 entities participating in the Sophia beta test at the end of the test period, December 31, 2012.

2.2.2 Path Forward

Sophia has been handed off to INL Technology Deployment (TD) for licensee evaluation. TD is working with about 10 potential licensee candidates to find the best plan for deploying Sophia into the SCADA market. Sophia is being submitted as a candidate for a R&D 100 Award in 2013.

3. MESH MAPPER PROJECT

The Mesh Mapper (MM) Tool was intended to passively collect the route information of a Wireless Mesh Network (WMN) and graphically present this information for quick analysis.

Complete details on the work done on the Mesh Mapper tool are provided in INL Control System Situational Awareness Technology Annual Reports for FY 2011 and FY 2012.

3.1 FY 2013 Q1 Mesh Mapper Update

There was no work done on the Mesh Mapper tool in FY 2013 Q1. A No-Go decision was made by the project team and DOE-OE CEDS PM after the FY 2012 Informal Design Review.

3.2 Summary

3.2.1 Accomplishments

The initial functional design requirements were completed at the beginning of FY 2011. The core functionality for a parsing engine for the C12.X protocol has successfully been developed. The parsing engine can currently do the following:

- Display unique Access Point-Titles (AP-titles) associated with a meter, cell relay, and the collection engine.
- Determine and display the node type of a given AP-title: meter, cell relay, or collection engine.
- Identify and display who has communicated with a given AP-title.
- Record and display the last time an AP-title communicated.
- Record and display the number of times an AP-title was either calling or being called.
- Determine and display the current Cell Relay (CR) that a meter is using.
- Calculate and display total number of known meters/devices on the network.
- Began development of a Java Graphical User Interface (GUI) front end and the design for integrating it with the parsing engine.

3.2.2 Path Forward

Based on the overall efforts of the MM project there is a need to research current mesh network monitoring capabilities and provide recommendations for enhancing them in the future. Research efforts should focus on:

- Existing vendor and third party monitoring capabilities and developments,
- Utility and asset owner desired capabilities for mesh network monitoring, current protocol capabilities and developments to support mesh monitoring,
- Analysis of the current mesh network standards.

The results of this research could be published in a white paper that details the findings and could provide recommendations for enhancing current and new standards that will be used by industry to improve mesh network monitoring capabilities.

4. INTELLIGENT CYBER SENSOR PROJECT

The Intelligent Cyber Sensor (ICS) looks at the expected traffic for each group of smart grid sensors and has a highly efficient mechanism of monitoring and filtering network performance data. When integrated with the University of Illinois NetAPT product, the ICS seeks to more effectively implement security policy, flag degradation trends in the associated sensor subsystem, and present information in an efficient, ergonomic fashion.

Complete details on the capabilities and progress of the ICS are provided in INL Control System Situational Awareness Technology Annual Report for FY 2011 and FY 2012.

4.1 FY 2013 Q1 Intelligent Cyber Sensor Update

ICS work in the first quarter of FY 2013 focused on two areas: 1) reducing false positives, and 2) implementation of commercial friendly features. Two distance measures used in the anomaly clustering algorithm were explored. The compression routines from the DOE-OE CEDS funded INL SCADA Protocol Anomaly Detection (SPADUC) project were incorporated as an additional anomaly data feature. Additionally, existing alerts were aggregated and syslog functionality was added to the sensor. Finally, modifications to the communication system were included to improve information passed for display by the DFT. While not all of the avenues proved fruitful, progress was made in the communication and alert aggregation functionality.

Technical details for Intelligent Cyber Sensor work completed in FY 2013 Q1 can be found in Appendix A.

4.2 Summary

4.2.1 Accomplishments

- Integrated SPADUC library compression routines as an additional data feature for anomaly detection.
- Examined the effectiveness of two alternate distance measures for use in the anomaly clustering routines.
- Implemented alert aggregation routine resulting in a 77 percent reduction in false positive alerts.
- Added a syslog capability for improved communication of alerts to existing operational components.
- Improved status and information communication mechanism with DataFusion HMI process.
- Submitted two revised drafts for IEEE Transaction special sessions.

4.2.2 Path Forward

The project recommendation is to move this concept to an Industry led project to seek a workable solution to deal with the remaining error rate in order to make this technology a viable product for use in a production utility environment.

5. DATA FUSION SYSTEM PROJECT

The Computational Intelligence (CI) Advanced Data Mining Techniques (ADMTs) combine multiple temporal/spatial, highly diverse data streams into a unified data model, then identify relationships and frequent data patterns that are common to SCADA and sensor systems, such as typify smart grids. As a result, resilient data fusion (generation of actionable intelligence) is achieved via various CI techniques. The Data Fusion Tool (DFT) is intended to couple the cyber health and operational performance aspects of a sensor network, including smart grid components, to provide an overall network performance indicator.

Complete details on the capabilities and progress of the DFT are provided in INL Control System Situational Awareness Technology Annual Report for FY 2011 and FY 2012.

5.1 FY 2013 Q1 Data Fusion Update

During the first quarter of FY2013, the Data Fusion Tool (DFT) development was continued primarily to enhance integration with the Intelligent Cyber Sensor (ICS). The cyber alert interface was modified to display a graph of the sensor activity including frequency of alerts and confidence threshold. User feedback mechanisms to change the operating mode or suggest changes to threshold to ICS were also included. The ability to plot the region of affected sensors has added to the Geographical Context Visualization Tool (GVCT). The GVCT is now utilized in multiple internal INL projects and additional users and future projects using it continue to be pursued.

Technical details for Data Fusion work completed in FY 2013 Q1 can be found in Appendix B.

5.2 Summary

5.2.1 Accomplishments

Team Completed demonstration of the DFT integrated with Sophia and ICS in September.

DFT progress is focused on the Geographic Context Visualization Tool (GCVT) and the integration with ICS and Sophia alerts through the IF-MAP publish/subscribe communication platform. Data prioritization methods have been applied to sensor data to determine correlated data feeds. This prioritization and correlation are captured in the visualization tool as with methods to show alternative data when a particular data feed is compromised as identified by a cyber alert or system malfunction. Integration of Situational Awareness Tools was accomplished through IF-MAP interface.

Collaborative efforts with Fujitsu Laboratories of America, Inc. (FLA) continued. Discussions on interest in visualization tool are ongoing. They have potential interest in the GCVT. The project continued to collaborate in forming value based metrics for health monitoring and potential optimization for energy systems.

Adaptive Critic and power models continue to develop and provide an additional virtual data feed to the data fusion tools. The work completed on this will be an asset for future INL and DOE projects. Two conference publications have been produced from collaborators at University of Idaho with funding credit to the project.

5.2.2 Path Forward

Potential for future use of the products of the project within INL and other applications are being pursued. Specifically, use for integrating a suite of resilient control system tools will carry on through Laboratory Directed Research and Development (LDRD) support through the Instrumentation, Control and Intelligent Systems (ICIS)[a] distinctive signature program. Some interest found through discussions with industry representatives at the integration demonstration will be pursued as potential use of the

[a] ICIS; https://inlportal.inl.gov/portal/server.pt/community/distinctive_signature__icis/315, viewed 1/30/2013

technology and possible future funding. Many potential paths forward exist for this platform. Additional features for driving the GCVT via “backroom” data mining to automatically bring the most important information to the forefront and provide a comfortable but attention stimulating display for operators are possible and should be pursued with the input of human factor engineering principles.

The DFT is adaptable to any analysis and prediction capability that the advanced data mining may provide. Any application requires analysis of the system for the “keystone” attributes for given operator, supervisor, and stakeholder contexts. This platform is readily adaptable to the needs of applications that require geographic context of the process information with a cyber aware need. Smart grids are just one such application for the GCVT data fusion tool. Higher-level decision advice can be provided by an adaptive critic mechanism through implementation of cost functions that provide guidance for demand response and utilization of dispatched local resources (e.g., energy generation or storage).

Appendix A

FY 2013 Q1 Intelligent Cyber Sensor Work Completed

ICS work in the first quarter of FY 2013 focused on two areas: 1) reducing false positives, and 2) implementation of commercial friendly features.

Research Items

1. Exploring possible improvements to the anomaly detection clustering algorithm.

The original algorithm was designed to work within the physical constraints of a smaller sensor device. There are several candidates we can explore within the current framework that may prove beneficial. The research team has done work in the past in this area with mixed results.

The Euclidean distance is a familiar geometric distance based on the Pythagorean formula. This distance measure is relatively simple to calculate using the following formula where ***x*** and ***y*** are *n* dimensional vectors representing points:

$$d(x,y)=\sum_{i=1}^{n}\sqrt{(x_i-y_i)^2} \qquad (1)$$

This distance measure has a straightforward geometric interpretation, is computationally inexpensive and simple to code; however, it does have two drawbacks. First, in geometric problems, domain variables are typically measured utilizing the same units of length. Data values from real world problems may have different scales. For example, a regression problem making use of class information such as age, test scores, and time are all on a different scale and therefore not directly comparable. The Euclidean distance is sensitive to the scales of the variables involved and may not perform optimally. This problem can be overcome by a standardized or weighted Euclidean distance that incorporates variance but not covariance. A Mahalanobis distance incorporates both variances and covariances.

Second, the Euclidean distance does not compensate for correlated variables. Given a test data set containing multiple variables where one variable set is an exact duplicate of another set, these sets are highly correlated. The Euclidean distance calculation will weight the duplicate variables more heavily than the others. It has no method of accounting for the fact that the duplicate provides no new information.

P.C. Mahalanobis introduced the Mahalanobis distance in 1936. It is based on both the mean and variance of the variables in addition to the covariance matrix. The iso-surface formed around the mean is an ellipse in two-dimensional space or an ellipsoid or hyper-ellipsoid when more variables are used. It is a multivariate quantitative method that can solve for multiple dimensions simultaneously. The covariance among the variables is taken into account when calculating the distance. Because of this, the problems of scale and correlation inherent in the Euclidean distance are not an issue. Given an individual as a vector $\vec{x}_i=(\mathrm{x}_0,\ldots,\mathrm{x}_n)$ of floating point values *x*, a vector representing the mean of a data set $\vec{\mu}=(\mu_0,...,\mu_n)$ and a covariance matrix C of size *n x n* representing the covariance values between all dimensions *n*, the Mahalanobis distance is calculated with the given formula:

$$md(\vec{x}_i)=(\vec{x}_i-\vec{\mu})C^{-1}(\vec{x}_i-\vec{\mu})^T \qquad (2)$$

This function produces a distance value for the input $\vec{x}_i$ vector.

According to formula two, three variables need to be defined. Trivially, the first is a feature vector x, which is gleaned from the input network packets. The second is a sample mean. The ICS code was changed to calculate and store the mean of the 17 network features found during the initial learning phase. The covariance matrix C is a matrix of all the covariances between each feature vector. Given a data point

and a sample mean, the covariance can be calculated. In the case of ICS, this produces a 17 by 17 covariance matrix. It should be noted that the formula specifies an inverted, or nonsingular, covariance matrix. Not all matrices are invertible. We calculated the Eigen decomposition of the sample covariance matrix. This produced Eigen values that were zero and thus indicating that the matrix was not invertible. This prohibited the use of Mahalanobis as a distance measure with the original ICS features.

It has been noted that the use of a Euclidean distance measure on high dimensional data can be problematic. When used as a nearest neighbor measure in high dimensions, data points that appear to be relatively close in low dimensions can be falsely determined in higher dimensions. One possible solution to this is to use a cosine similarity measure. This measure is the cosine of the angle between two vectors. Each data point is treated as a vector. The Euclidean dot product, or inner product, is calculated as in Figure 1 and equates to the cosine of the angles between the vectors. The angle thus determines whether two vectors are pointing in roughly the same direction.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Figure 1. Cosine Similarity.

The unmodified ICS anomaly detection routine utilizes a 17 dimensional feature set. Each data point is normalized to a floating point value between 0.0 and 1. The cosine similarity algorithm was applied to a test set of network data. The ICS algorithm utilizes the distance measure as a key criteria for a binary decisions. The first condition is to determine if a candidate point falls within a predefined configurable distance to already defined clusters. If this occurs, the point is added to the cluster with the minimum distance. If not, then a new cluster is created and the candidate point is the exemplary member. For our normalized Euclidean distance, this value was 0.25. The definition of this value for a cosine angle was explored. We were unable to find a single value that when applied to all the data could adequately separate the anomalies from normal vectors. There were individual cases of Internet protocol (IP) addresses where a defined value would differentiate between IP's. However the accuracy never surpassed that of the original algorithm. It is surmised that the number of data points in the feature points made the differences in the angle too small to detect. This is partially supported by the effectiveness of the original Euclidean measure based system. In a test run, the cosine distance measures were captured, this resulted in 23,021,335 entries. The run produced 37,6166 unique values with a minimum value of 0.587196 and a maximum of 1. The average of the values was 0.81099 with a standard deviation of 0.11011.

Another concept explored with the clustering algorithm was changing the cluster comparison points. The original algorithm computes the distance of a candidate point to a representative center point of the existing clusters. Upon acquiring a new data vector $\vec{x}_i$ from the shifting window buffer, the set of clusters is updated according to a modified online Nearest Neighbor Clustering algorithm. First, the Euclidean distance to all available clusters with respect to the new input feature vector $\vec{x}_i$ is calculated. The nearest cluster P_a is identified. If the computed nearest distance is greater than the established maximum cluster radius parameter, a new cluster is created. Otherwise the nearest cluster P_a is updated according to the following:

$$\vec{c}_a = \frac{w_a \vec{c}_a + \vec{x}_i}{w_a + 1}, \; w_a = w_a + 1 \quad (3)$$

$$c_{i,j}^U = \max(x_i^j, c_{i,j}^U), \; c_{i,j}^L = \min(x_i^j, c_{i,j}^L) \quad j = 1 \ldots n \quad (4)$$

In our proposed modification, instead of keeping a weighted representative point calculated using 3 and 4, an actual cluster point was chosen as the cluster central point. This modification was run on test data and the results are shown in Table 1 and Table 2. Compared to the original results the modification did not achieve improved performance.

IP Address	Correct	Incorrect	False Negative	False Positive
*.99.5	99.9	0.1	0.11	0.001
*.99.206	99.99	0.01	0.01	0.06
*.99.101	99.97	0.03	0.02	0.0

Table 1. Original Results

IP Address	Correct	Incorrect	False Negative	False Positive
*.99.5	99.87	0.13	0.58	0.0
*.99.206	99.69	0.31	0.44	0.04
*.99.101	99.8	0.2	0.50	0.0

Table 2. Modified Results

2. New anomaly data feature.

The current ICS inspects network packet data. The two classes of information include 1) statistics and information about packets that reside within a packet buffer, i.e., time differences, number of protocols seen, and 2) individual packet header information such as protocol type, payload size, etc. The information contained within the network packets is not being fully exploited. Features derived from the packet payloads may improve accuracy and precision. A potential source of packet payload information can be found in the project.

Initially, to get a baseline for the use of compression in ICS, the commonly used zlib compression library was integrated. The zlib routines report enough information to determine the resulting byte size of a compressed data buffer. This value was added as the only feature instead of the existing 17 feature set. Two data compression candidates are individual network packet data portions and a buffer filled with all 20 network packet data portions concatenated together. The ICS algorithm uses a sliding 20 packet buffer routine to calculate inter packet statistics. A test using both data candidates was performed on a single IP. The results of the two runs are shown in Table 3. With the exception of false negatives the buffer approach provided better results. This implementation was chosen for further exploration.

Data Source	Correct	Incorrect	False Negative	False Positive
Avg. Single Packet	78.49	21.51	45.3	14.6
20 Packet Buffer	83.05	16.95	75.2	0.02

Table 3. Initial Compression Results

Utilizing the compression library resulting from early SPADUC project results, an implementation using the 20 packet buffer was explored. The SPADUC library routines return the number of encodes performed on the data buffer. All the standard ICS features were removed and only the SPADUC encode value was considered. The results of this testing on nine different IP network streams are shown in Table 4. A general trend observed is a high false positive rate and a low false negative rate. The variation in percentage values might be attributable to the heterogeneous nature of the devices tested. Given no other data than the number of encodes, this can be interpreted as a good possibility to add as a feature for anomaly classification.

IP Address	Correct	Incorrect	False Negative	False Positive
*.99.5	91.68	8.3	0	53.71
*.99.101	99.67	0.33	0	3.46
*.99.107	95.32	4.68	0	26.91
*.99.135	90.55	9.45	0	100
*.99.140	96.98	3.02	0	62
*.99.160	35.49	64.5	26.7	72
*.99.170	68.99	31	0	99.9
*.99.206	97.60	2.40	0	46.8
*.99.220	21.05	78.95	0	94.49

Table 4. SPADUC compression feature

3. Alert aggregation and syslog capability

A comment made to the Sophia developers during the industry presentation stated that syslog is still a widely used distributed information mechanism. Adding this capability to the alerts functionality would improve its integration possibilities with COTs type software.

Anomaly alerts occur for each packet that is identified. This makes for a possibly large amount of alerts depending upon the traffic seen. We have observed that the anomalous traffic tends to occur in bursts of related traffic. A mechanism was implemented to aggregate related alerts together and to result in a single alert. This has several desirable effects. The first is the reduction of alert traffic for the same incident. The second is the possible reduction of false positives. The false positive traffic has also been observed to occur in clusters. This type of mechanism would not repair the recognition mechanism but would effectively filter the false positive reporting and reduce the volume sent to a user.

Log::Log4perl is a Perl port of the popular log4j logging package. It allows for controlling the amount of logging messages generated very effectively. Logging levels can be changed during the running of a process. Additionally messages can be redirected to multiple outputs such as a file, email, or database. This occurs without having to changing a program's source code. The ICS utilizes a standard centralized messaging interface based on PERL. Therefore Log4perl is a natural candidate to implement both syslog capability and managing alert message volume.

Log4perl has a concept called appenders. Appenders define logic applied to log data to be written to output devices. There are default appenders provided as part of the distribution, but custom logic can be implemented as well. This later functionality was utilized by the ICS project. A custom appender was developed to aggregate Alerts from the Anomaly detection routine. The appender will consume an alert as it is created and store the information in an internal buffer. Any subsequent alerts are then examined for similarity to the buffered alert. If it is identical in IP addresses, port numbers, and IP protocol, then a counter value is incremented. If these criteria are not met, then the buffered alert is released and the current alert is buffered. This has the effect of reducing the amount of alerts sent from the Sensor device. In addition to a change in alert values if the incremental counter reaches a configurable limit the message will be sent. A trivial enhancement to the appender would be to add a default time limit on how long alerts are held. This would ensure timely delivery of alerts if the monitored network has few anomalies. In addition to the aggregation logic, a syslog output capability was added.

The test system, when run initially without aggregation, produced 178,191 alerts to the system syslog file. Of these alerts, 1,549 messages were false positives. The aggregate appender was enabled with an increment count maximum of 20. The total number of alerts was reduced to 29,270 and the false positive message totaled 845. These messages included the criteria mentioned previously with the addition of a confidence value. The 20 count maximum test was run without considering the confidence value in the

message similarity logic. The total number of alerts from this change was 17,631. The number of false positives was reduced to 352. These simple measures reduced the false positive rate to 23 percent of the original. An example message is shown next. The 12 in brackets is the value of the increment counter for the message.

2012/12/10 13:31:33 INFO [12]: CS:ID1:001:IAA:1:anomaly:20:192.168.99.135 192.168.99.10 192.168.99.206 192.168.99.140:6:43.52124 -112.05260

4. HMI work.

The sensor has always been a standalone component with information provided by a messaging system. It was designed to work with other components and does not have a control or management GUI. Some work with the Data Fusion GUI has been done. In order to improve ICS for standalone use, we expanded the ICS messaging and Data Fusion HMI capabilities.

ICS Integration into Situational Awareness

One method in which ICS integrates into the overall situational awareness architecture is through alerts and status provided through the IF-MAP Message Interface. Alerts contain information to provide cyber-physical operation experts the tools to evaluate the information available from ICS. The data feeds include mode status (train, monitor, online, offline, etc), confidence in anomaly detection, adaptive threshold for alerts, and alert status. Figure 2 is a bock diagram of the interconnections between components, showing the display containing the detailed history of the ICS as presented by the Data Fusion Tool (DFT).

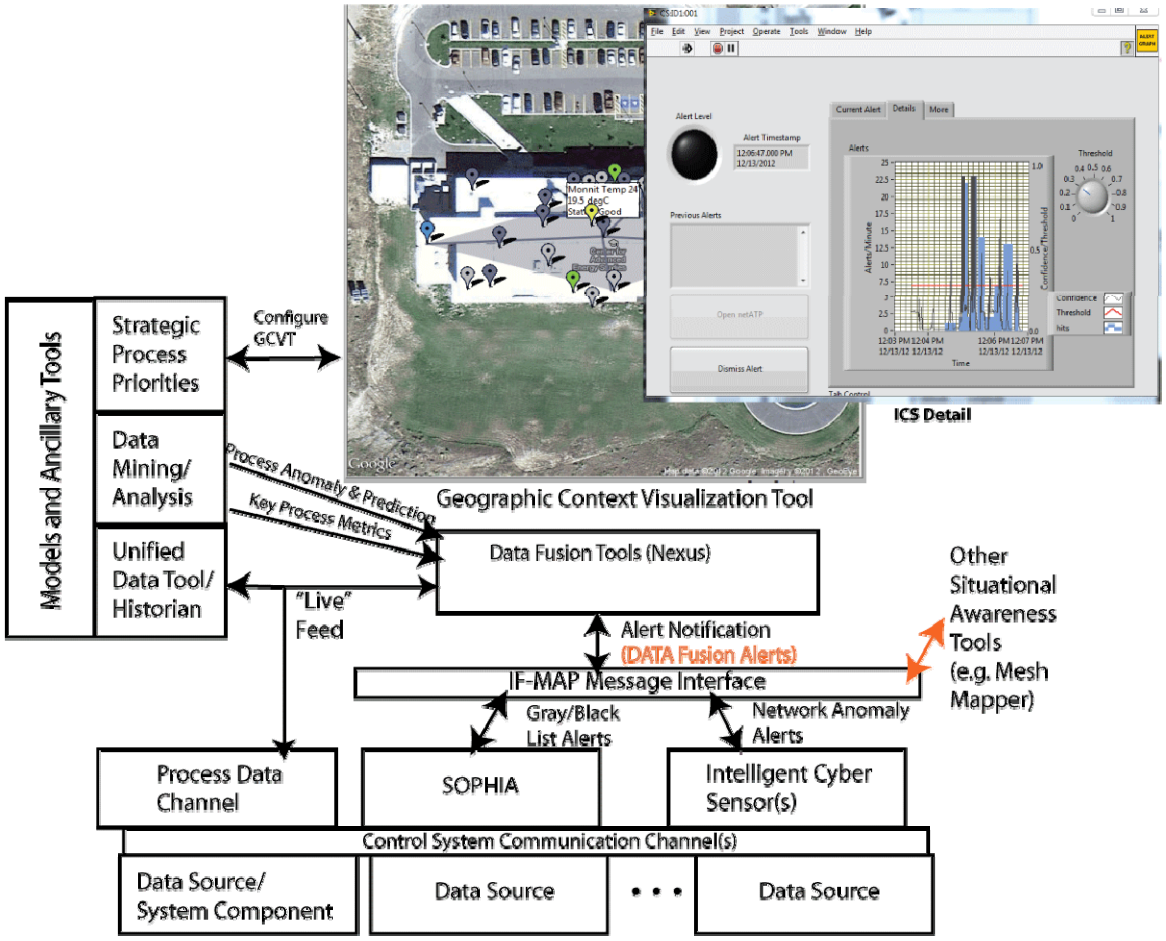

Figure 2. Overview of integration of ICS into Situational Awareness Architecture

The messaging from ICS is in XML and is parsed and aggregated by the DFT. The details of the message are as follows:

<resultItem>
- <device>
<name>CS:ID1:001</name>
</device>
- <metadata>
- <meta:event xmlns:meta=”http://www.trustedcomputinggroup.org/2010/IFMAP-METADATA/2” ifmap-cardinality=”13ultivalued” ifmap-publisher-id=”sensor-1904628644-1” ifmap-timestamp=”2012-12-11T13:40:05-07:00”>
<name>IAA</name>
<discovered-time>2012-12-11T20:40:18</discovered-time>
<confidence>0.117505</confidence>
<type>normal [/anomaly]</type>
<window-size>20</window-size>
<threshold>0.3</threshold>
- <ip_list>
<ip>192.168.99.135</ip>
<ip>192.168.99.206</ip>
<ip>192.168.99.140</ip>
</ip_list>
- <proto_list>
<proto>6</proto>
</proto_list>
- <position>
- <Point srsDimension=”2”>
<pos>43.52124 -112.05260</pos>
</Point>
</position>
</meta:event>
</resultitem>

From the message, the DFT can aggregate time versus confidence and threshold as well as the number of alerts. This history is displayed in the details of the ICS information in a child window of the DFT that is drilled down to by clicking on the marker location of the ICS, which will alternate color when an unacknowledged alert has occurred. Figure 3 shows the representation of the time series history from the ICS derived from the ICS messages. In this display, the user may provide an elevated local threshold for alerts. For example, the operator may not be qualified to evaluate this specialized information and may not want the alerts. A message informing the system of the user set threshold may be sent to the IFMAP in a future version of the software to inform the system that the operator is not viewing every alert. This enables system situational awareness of operator interactions with the system. Currently there is not a supervisory component of the system to “pay attention” to such interaction but this information could be tapped by an existing component of an application or a supervisory role enhancement to the DFT.

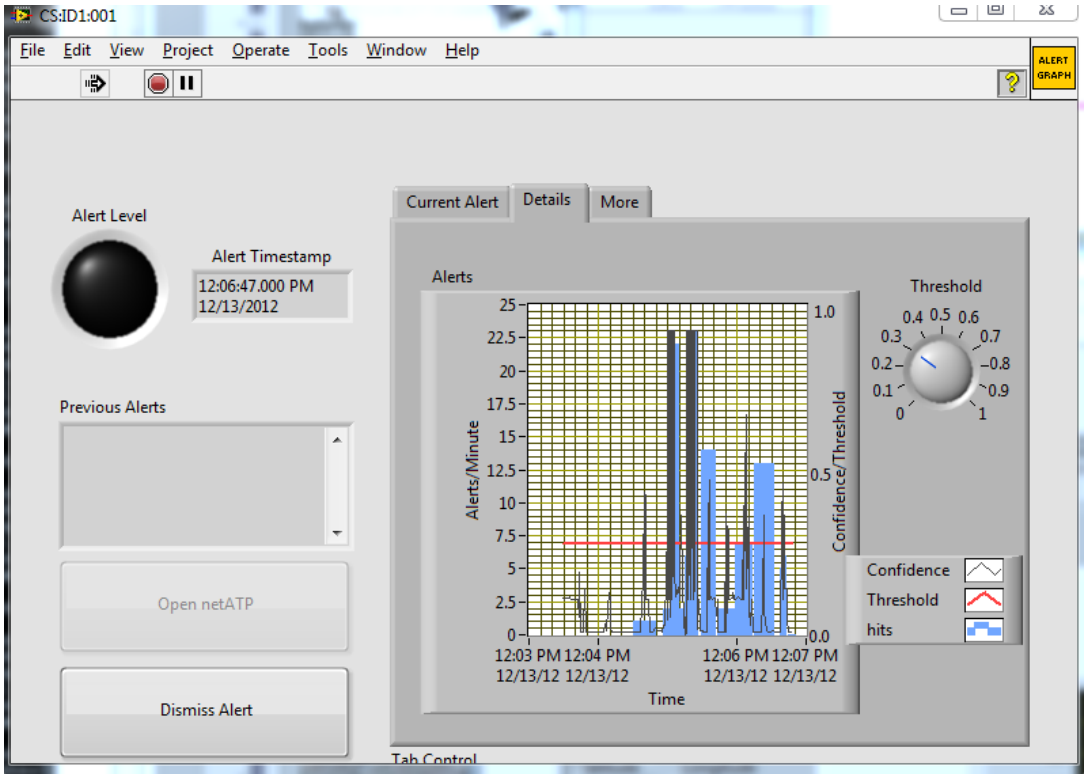

Figure 3. Operator sees the historical confidence level of anomaly detection as well as the number of alerts and threshold

Geographic connection of the ICS message is obtained from the <pos> tag. The DFT uses this to generate the location of the marker. Additionally, when the IP addresses can be associated with data sources or other markers, the DFT is aware of a "web" graphic that will be generated when the operator selects the ICS device location. Figure 4 shows an example of a device that is associated with other devices with visualization through generation of an encoded path command to google maps. If IP address associated with the alert are common to a known device location, a mapping of the affected elements is rendered through the connecting lines and shaded region. An operator therefore has a geographical context for the alarm that can aid in decisions about the overall distributed system. For example, data may be given a confidence or weighting in a decision process in a dynamic manner. Weight on those parameters in an optimizing algorithm can be attenuated.

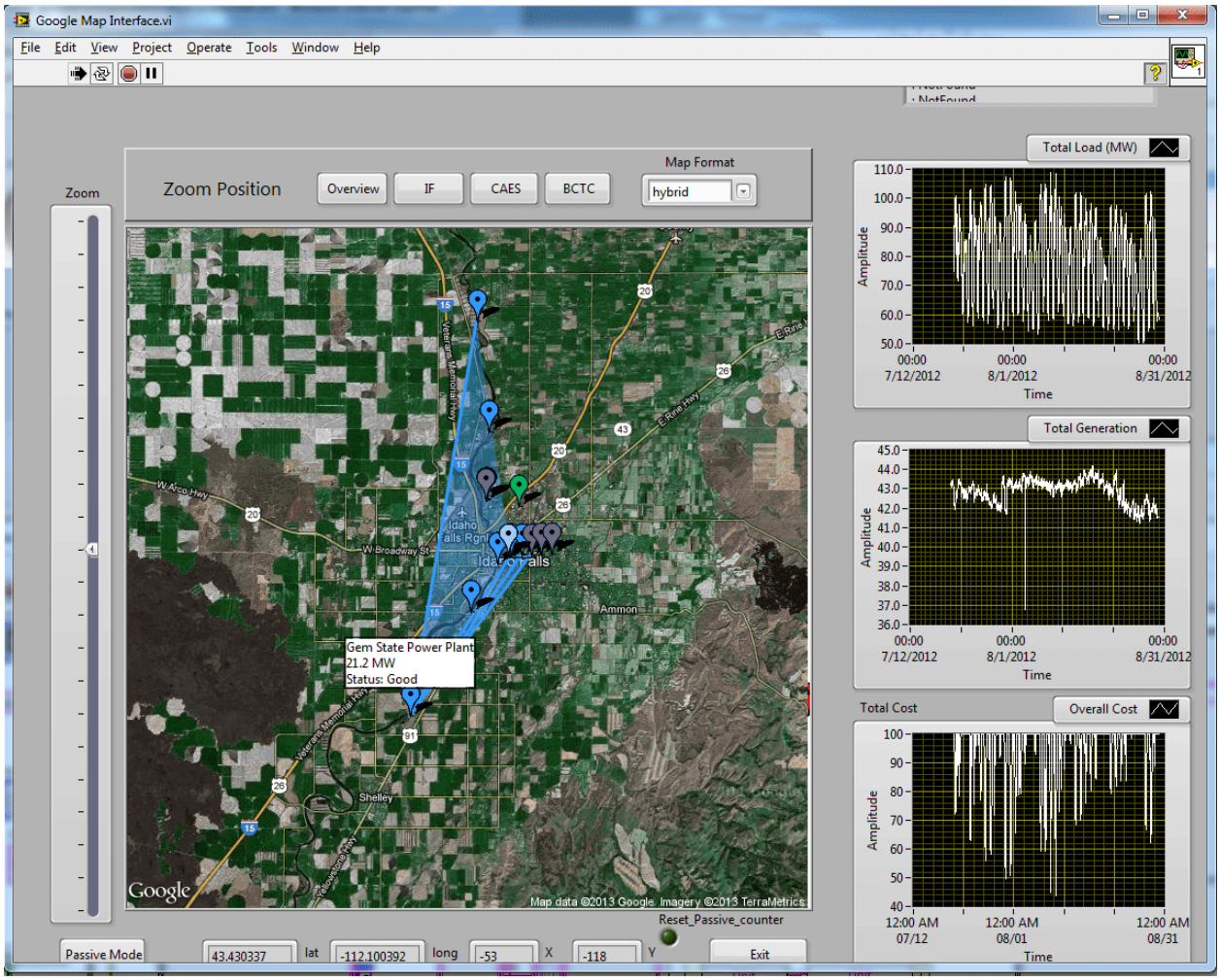

Figure 4. Example of visualization of sources related to a sensor or cyber event based on associated IP Addresses

The operator viewing the DFT display has the ability to configure the ICS mode through a drop down menu, where the available modes are displayed and selectable, as shown in Figure 5.

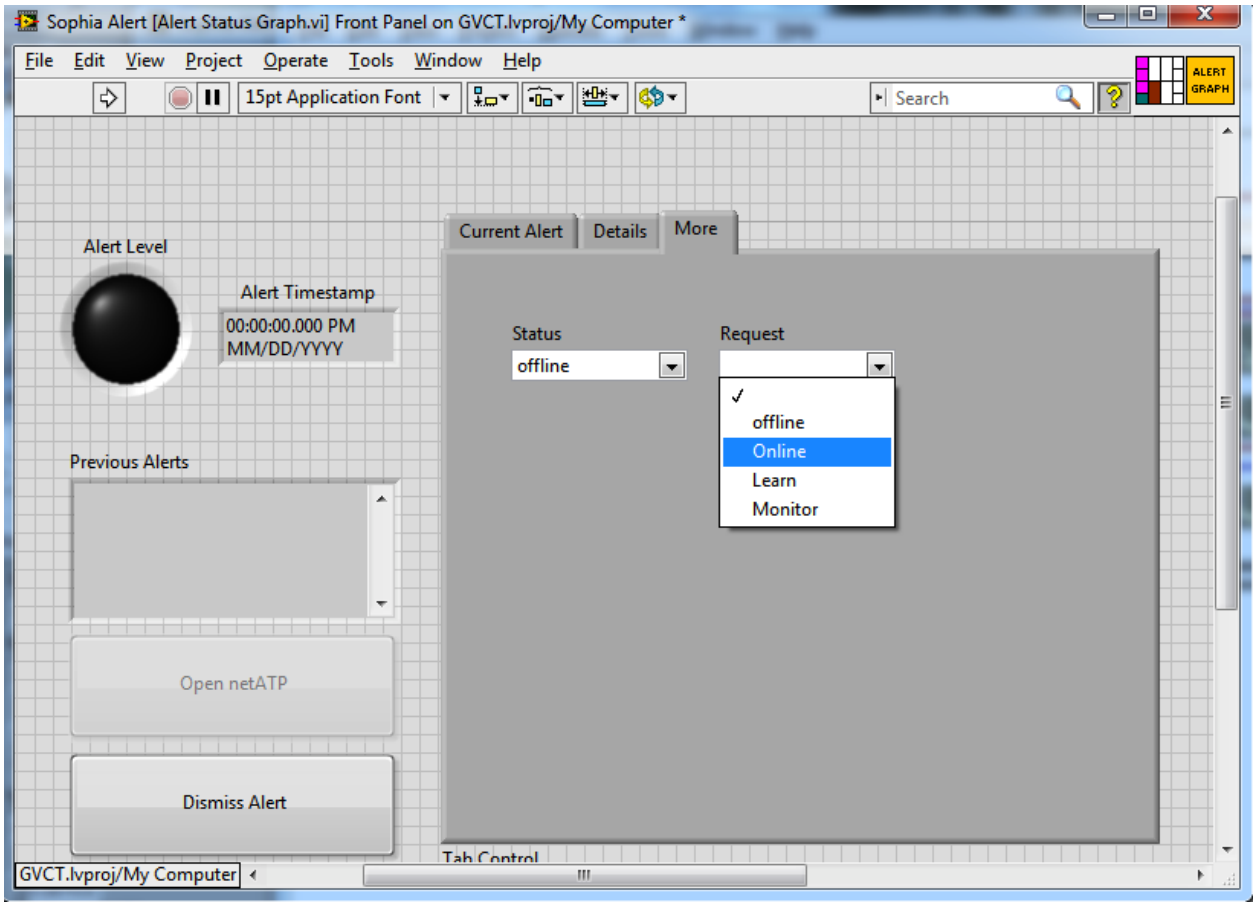

Figure 5. Operator configuration of the ICS mode through the DFT interface

Details of that message are simply the sending device followed by the targeted receiver (i.e., ICS) and the request for change in mode. The XML structure sent through the IF-MAP interface is where the boldfaced text will vary based on request:

<resultItem>
<device><name>DF:ID1:001</name></device>
<device><name>CS:ID1:001</name></device>
<metadata><meta:sensor xmlns:meta="http://www.trustedcomputinggroup.org/2010/IFMAP-METADATA/2" ifmap-cardinality="singleValue" ifmap-publisher-id="visual--1781274946-1" ifmap-timestamp="2013-01-02T17:34:45-07:00">
<request>**offline**</request>
</meta:sensor>
</metadata>
</resultItem>

The ICS device will respond similarly to indicate the mode has been changed. Other messaging that becomes of use for ICS or DFT will be defined at a later time if/when development continues.

ICS/DFT Future Direction

There is opportunity in the future for integration of the situational awareness tools to be enhanced. Specifically, the DFT can become a more complete front end to the ICS solutions. The vector produced from anomalies could be further processed offline and analyzed offline. Through operator interaction, the thresholds may be adapted to provide more desirable constraints on alerts by incorporating it as a training

variable in the fuzzy-2 algorithms. Additionally, the vector of the anomaly is a “description” of the pattern of specific bad actors. Offline these can be clustered and classified in comparison to known signatures. The ICS can also be configured to archive more extensive data sets based on user driven requests. As an example, a set of anomalies that occur on occasion could be declared of high interest and entire network packets that produce a signature close to this can be stored for more extensive forensics, thereby achieving the ability to identify new potential threats without the need for intractable processing of entire network streams.

Appendix B

FY 2013 Q1 Data Fusion Work Completed

During the first quarter of FY 2013, the Data Fusion Tool (DFT) development was continued primarily to enhance integration with the Intelligent Cyber Sensor (ICS).

GCVT operator interaction with ICS.

Advanced tabs have been added to the Geographical Context Visualization Tool (GCVT) Graphical User Interface (GUI) providing users the opportunity to review the state and training process of the ICS; specifically, the history of confidence level, the number of alarms in a time period, and the adaptively configured fuzzy threshold for alarms. Alarms are indicated with a geographic location. Figure 6 shows the generation of a new marker accompanied by an automatic display zoom to a level and position that leaves both the previous focus and the new information in context. The marker location of the active alert will flash at a 1 hz frequency until an operator responds.

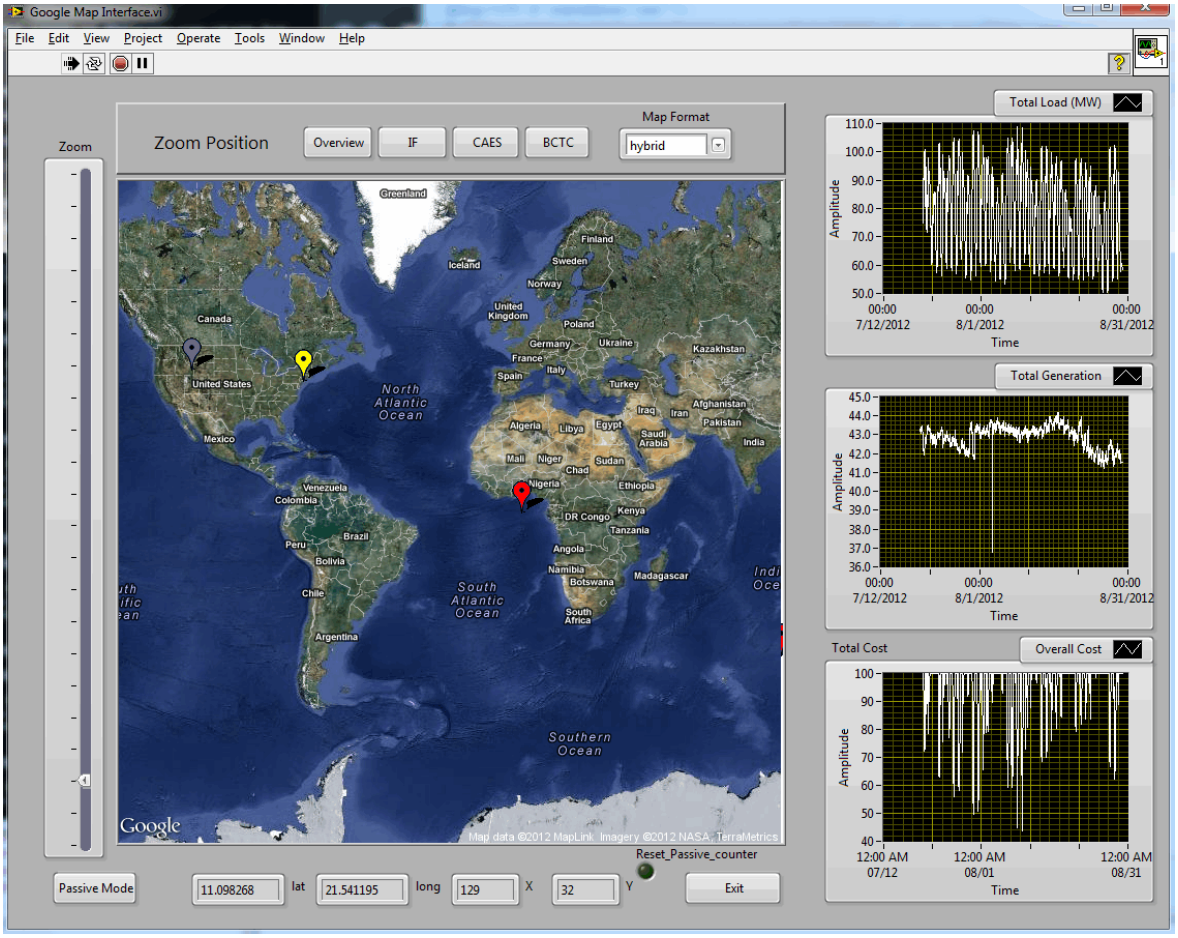

Figure 6. New communication alarms are automatically generated with automatic zoom function to put current and new context into view

Operators with expertise in the ICS may now see details of the performance of an ICS as shown in Figure 7. The number of alerts per minute and the historical anomaly confidence factor and threshold are available for evaluation. A local threshold may be set for this user interface to a level higher than the automatic set to decrease the frequency of alerts. This operator adjusted parameter is fed back to the ICS as a potential future variable to incorporating expert interaction into the algorithm.

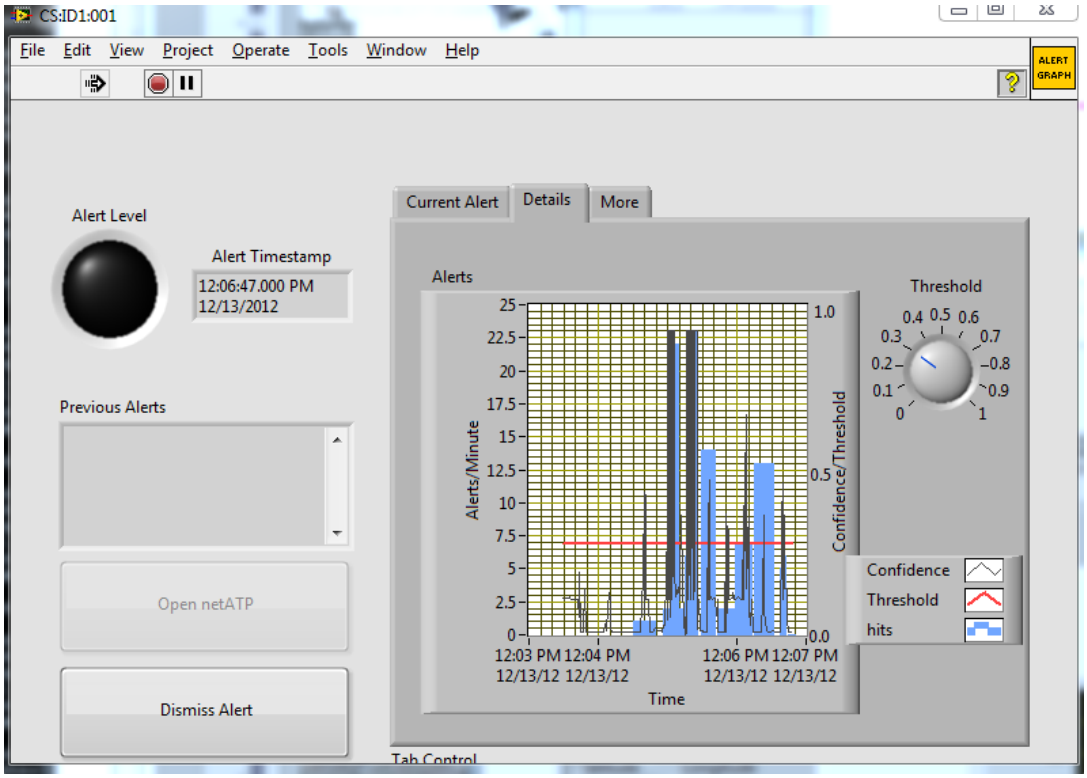

Figure 7. Operator sees the historical confidence level of anomaly detection as well as the number of alerts and threshold

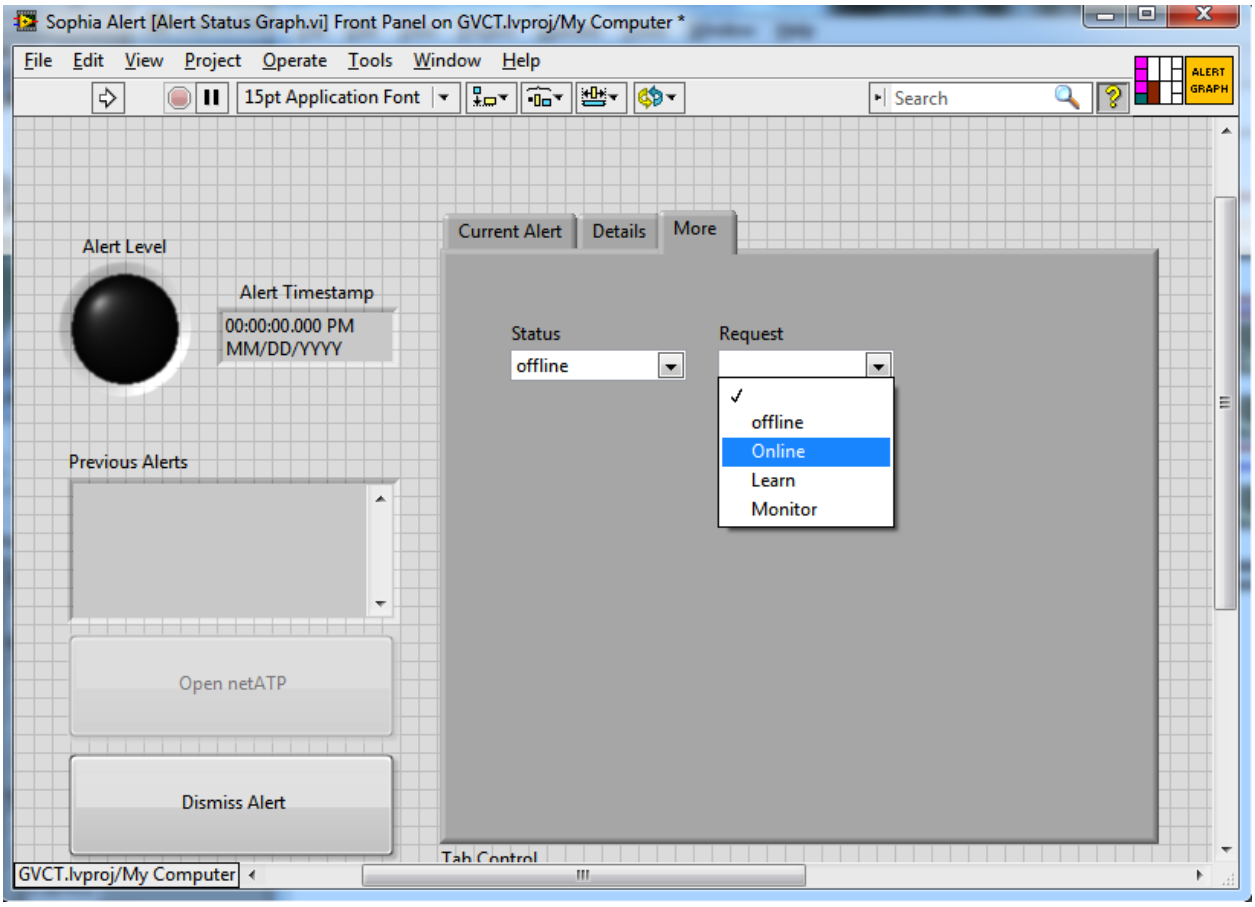

Figure 8. The operator may configure the mode of operation of ICS through the interface

Additional messaging to the ICS may be initiated through the interface. An operator selecting one of the modes will produce a message to the IFMAP interface (see Figure 8). The ICS will respond with a status message when the mode is successfully changed. Other opportunities for configuration of ICS may be enabled through this interface in the future. This control may be used in concert with GCVT or as a standalone interface to the ICS.