First International Diagnosis Competition – DXC’09

Tolga Kurtoglu*, Sriram Narasimhan**, Scott Poll***, David Garcia****, Lukas Kuhn†,
Johan de Kleer‡, Arjan van Gemund¶, Alexander Feldman**

* Mission Critical Technologies @ NASA Ames Research Center
** University of California, Santa Cruz @ NASA Ames Research Center
*** NASA Ames Research Center
**** Stinger Ghaffarian Technologies @ NASA Ames Research Center
† Palo Alto Research Center
‡ Delft University of Technology

Abstract: A framework to compare and evaluate diagnosis algorithms (DAs) has been created jointly by NASA Ames Research Center and PARC. In this paper, we present the first concrete implementation of this framework as a competition called DXC’09. The goal of this competition was to evaluate and compare DAs in a common platform and to determine a winner based on diagnosis results. 12 DAs (model-based and otherwise) competed in this first year of the competition in 3 tracks that included industrial and synthetic systems. Specifically, the participants provided algorithms that communicated with the run-time architecture to receive scenario data and return diagnostic results. These algorithms were run on extended scenario data sets (different from sample set) to compute a set of pre-defined metrics. A ranking scheme based on weighted metrics was used to declare winners. This paper presents the systems used in DXC’09, description of faults and data sets, a listing of participating DAs, the metrics and results computed from running the DAs, and a superficial analysis of the results.

1. INTRODUCTION

The DX community meets every year at the International Workshop on the Principles of Diagnosis1 to discuss the latest developments in the field of model-based diagnosis. Various diagnostic modeling approaches, associated reasoning algorithms, and applications to real and toy systems are presented. However, efforts to compare and evaluate diagnosis algorithms (DAs) on a common platform have been far and few in between. Other diagnosis communities have also not been actively involved in creating such platforms.

There have been attempts to benchmark DAs by computing performance metrics/indices (Orsagh et al., 2002, Bartys et al., 2006, Simon, et al., 2008). However these approaches were focused on specific domains and lack a general-purpose representation. In an effort to bridge this gap a framework called the DXC framework (Kurtoglu et al., 2009) was created to provide a level playing field to evaluate and compare DAs. This framework tried to establish a general purpose representation for system description, scenario data format, and diagnostic result format. A run-time architecture was created to execute DAs under similar conditions and compute performance metrics based on diagnostic output and ground-truth data.

In this paper, we present the first concrete implementation of the DXC framework called the DX Competition2 (DXC’09). 12 DAs (model-based and otherwise) competed in 3 tracks that included industrial and synthetic systems. Initially the participants were provided with system descriptions and a sample data set that included nominal and fault scenarios. The participants had to provide algorithms that communicated with the run-time architecture to receive scenario data and return diagnostic results. For the competition we ran all the algorithms on extended scenario data sets (different from sample set) to compute a set of pre-defined metrics. The metrics (with associated weighting) were used to rank the DAs.

The rest of this paper details the constituent pieces of this competition. Section 2 describes the tracks and systems used in the competition. Section 3 lists the classes of faults that were injected, how they were injected, and the sensor data sets that were generated as a result. Section 4 defines all the metrics used in the competition. Section 5 describes the conditions of the competition including information on how the algorithms were started and executed. Section 6 lists and briefly describes the participating DAs. Section 7 provides results of the competition and an analysis of the performance of the DAs. Section 8 introduces the assumptions that were made in implementing this competition, issues that were identified, and scope for possible extensions. Section 9 presents the conclusion and looks forward to the continuation of this competition in future years.

2. TRACKS & SYSTEMS

One of the primary goals of DXC’09 is to facilitate the development of domain independent diagnostic software. Furthermore, diagnostic software should be stress tested with difficult cases to determine its strengths and weaknesses and to pose a challenge. A diagnostic problem may be interesting due to its practical importance or it may be challenging due to

---

1 http://www.isy.liu.se/dx09/  
2 http://dx-competition.org/
its size and complexity. To facilitate all this we included multiple DXC tracks, and (optionally) multiple tiers in each track. The DXC'09 tracks and tiers are summarized in Table 1.

Table 1. Tracks, systems, and tiers in DXC'09

<table>
<thead>
<tr>
<th>Track</th>
<th>Tier</th>
<th>Systems</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industrial</td>
<td>1</td>
<td>ADAPT-Lite</td>
<td>Basic faults injected into a simplified EPS (Electrical Power System) tested</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>ADAPT</td>
<td>More complex faults injected into the full EPS distribution system</td>
</tr>
<tr>
<td>Synthetic</td>
<td>1</td>
<td>ISCAS85</td>
<td>Multiple faults injected into the circuits from the ISCAS85 benchmarks</td>
</tr>
</tbody>
</table>

2.1 Industrial Track

The hardware system for the DXC’09 Industrial Track is the Electrical Power System testbed in the ADAPT lab at NASA Ames Research Center (Poll et al., 2007). The ADAPT EPS testbed provides a means for evaluating diagnostic algorithms through the controlled insertion of faults in repeatable failure scenarios. The EPS testbed incorporates low-cost commercial off-the-shelf (COTS) components connected in a system topology that provides the functions typical of aerospace vehicle electrical power systems: energy conversion /generation (battery chargers), energy storage (three sets of lead-acid batteries), power distribution (two inverters, several relays, circuit breakers, and loads) and power management (command, control, and data acquisition). The EPS delivers AC (Alternating Current) and DC (Direct Current) power to loads, which in an aerospace vehicle could include subsystems such as the avionics, propulsion, life support, environmental controls, and science payloads. A data acquisition and control system commands the testbed into different configurations and records data from sensors that measure system variables such as voltages, currents, temperatures, and switch positions.

The scope of the ADAPT EPS testbed used for DXC Industrial Track is shown in Fig. 1. Tier 1 has the reduced scope as indicated. The nomenclature in the figure is consistent with the system description provided to all participants, which provides component, connection, and mode information. The characteristics of Tier 1 and Tier 2 are summarized in Table 2. The greatest simplification of Tier 1 relative to Tier 2 is not the reduced size of the domain but the elimination of nominal mode transitions. The starting configuration for Tier 1 data has all relays and circuit breakers closed and no nominal mode changes are commanded during the scenarios. Hence, any noticeable changes in sensor values may be correctly attributed to faults injected into the scenarios. By contrast, the initial configuration for Tier 2 data has all relays open and nominal mode changes are commanded during the scenarios. The commanded configuration changes result in adjustments to sensor values as well as transients which are nominal and not indicative of injected faults.

2.2 Synthetic Track

For the synthetic track, we have used the well-known benchmark models of ISCAS85 circuits (Brigelz and Fujiwara, 1985). These circuits are purely combinational, i.e., they contain no flip-flops or other memory elements. Note that the high-level structure of the ISCAS85 circuits, which can be beneficial to Model-Based Diagnosis (MBD) analysis, has been flattened out. A reverse engineering effort had resulted in high-level Verilog models (Hansen et al., 1999). Table 3 summarizes the circuits used in the synthetic DXC’09 track. Note that for many tasks of MBD (e.g., computing MFMC (Max-Fault Min-Cardinality) observations (Feldman et al., 2008)), the number of components in the ISCAS85 circuits can be reduced by performing cone identification (Siddiqi and Huang, 2007, de Kleer, 2008). The number of components in the reduced circuits is shown in the rightmost column of Table 3. We have left the decision if to identify cones to the competitors, i.e., we distribute the non-reduced circuits. In this first year of the competition we injected the complete fault at one instant. For example, if 3 components are faulted, the first observation provided to the DA is the result of all three faults injected simultaneously.

3. FAULT INJECTION AND SCENARIOS

3.1 Industrial Track

ADAPT supports the repeatable injection of faults into the system in one of three ways:

Table 2. Industrial track tier characteristics

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Tier 1</th>
<th>Tier 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Comps/Modes</td>
<td>37 / 93</td>
<td>173 / 430</td>
</tr>
<tr>
<td>Initial State</td>
<td>Relays closed; circuit breakers closed</td>
<td>Relays open; circuit breakers closed</td>
</tr>
<tr>
<td>Nominal mode changes?</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 3. ISCAS85 models (V and C denote the total number of variables and clauses, respectively)

<table>
<thead>
<tr>
<th></th>
<th>original</th>
<th>reduced</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sys</td>
<td>IN</td>
<td>OUT</td>
<td>IN/OUT</td>
</tr>
<tr>
<td>74182</td>
<td>9</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>74L85</td>
<td>11</td>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td>74283</td>
<td>9</td>
<td>5</td>
<td>36</td>
</tr>
<tr>
<td>74181</td>
<td>14</td>
<td>8</td>
<td>65</td>
</tr>
<tr>
<td>432</td>
<td>36</td>
<td>7</td>
<td>160</td>
</tr>
<tr>
<td>499</td>
<td>41</td>
<td>32</td>
<td>202</td>
</tr>
<tr>
<td>880</td>
<td>60</td>
<td>26</td>
<td>383</td>
</tr>
<tr>
<td>1355</td>
<td>41</td>
<td>32</td>
<td>546</td>
</tr>
<tr>
<td>1908</td>
<td>33</td>
<td>25</td>
<td>880</td>
</tr>
<tr>
<td>2670</td>
<td>233</td>
<td>140</td>
<td>1193</td>
</tr>
<tr>
<td>3540</td>
<td>50</td>
<td>22</td>
<td>1669</td>
</tr>
<tr>
<td>5315</td>
<td>178</td>
<td>123</td>
<td>2307</td>
</tr>
<tr>
<td>2688</td>
<td>32</td>
<td>32</td>
<td>2416</td>
</tr>
<tr>
<td>7552</td>
<td>207</td>
<td>108</td>
<td>3512</td>
</tr>
</tbody>
</table>

3. FAULT INJECTION AND SCENARIOS

3.1 Industrial Track

ADAPT supports the repeatable injection of faults into the system in one of three ways:
Hardware-Induced Faults: These faults are physically injected at the testbed hardware. A simple example is tripping a circuit breaker using the manual throw bars. Another is using the power toggle switch to turn off the inverter. Faults may also be introduced in the loads attached to the EPS. For example, the valve can be closed slightly to vary the back pressure on the pump and reduce the flow rate.

Software-Induced Faults: In addition to fault injection through hardware, faults may be introduced via software. Software fault injection includes one or more of the following: 1) sending commands to the testbed that were not intended for nominal operations; 2) blocking commands sent to the testbed; and 3) altering the testbed sensor data.

Real Faults: In addition the aforementioned two methods, real faults may be injected into the system by using actual faulty components. A simple example includes a blown light bulb. This method of fault injection was not used in the first DX Competition.

In addition, the software architecture described in (Kurtoglu et al., 2009) allows the injection of multiple faults into the system. Distinct faults types that are injected into the testbed for the DX Competition are shown Table 4 and summarized in Table 5.

### Table 4. Fault types used for the industrial tracks of DXC'09

<table>
<thead>
<tr>
<th>Component</th>
<th>Fault Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Battery</td>
<td>Degraded</td>
</tr>
<tr>
<td>Boolean Sensor</td>
<td>Stuck at Value</td>
</tr>
<tr>
<td>Circuit Breaker</td>
<td>Tripped, Failed Open, Stuck Closed</td>
</tr>
<tr>
<td>Inverter</td>
<td>Failed Off</td>
</tr>
<tr>
<td>Relay</td>
<td>Stuck Open, Stuck Closed</td>
</tr>
<tr>
<td>Sensor</td>
<td>Stuck at Value, Offset</td>
</tr>
<tr>
<td>Pump (Load)</td>
<td>Flow Blocked, Failed Off</td>
</tr>
<tr>
<td>Fan (Load)</td>
<td>Over Speed, Under Speed, Failed Off</td>
</tr>
<tr>
<td>Light Bulb (Load)</td>
<td>Failed Off</td>
</tr>
</tbody>
</table>

As shown in Table 5, nominal scenarios comprise roughly half of the Tier 1 and one-third of the Tier 2 competition scenarios. The Tier 1 fault scenarios are limited to single faults. Half of the Tier 2 faults scenarios are single faults; the others are double or triple faults. For both tiers once faults are injected they persist until the end of the scenario. In the case of multiple faults, they may be injected simultaneously or sequentially. In the first year of the competition the fault types are limited to additive parametric (abrupt changes in parameter values) and discrete (unexpected changes in system state).

### Table 5. Number of sample and competition scenarios for industrial track

<table>
<thead>
<tr>
<th>Scenario Type</th>
<th>Sample</th>
<th>Competition</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Scenarios</td>
<td>Tier 1</td>
<td>Tier 2</td>
</tr>
<tr>
<td>Nominal</td>
<td>32</td>
<td>39</td>
</tr>
<tr>
<td>Single-fault</td>
<td>27</td>
<td>54</td>
</tr>
<tr>
<td>Double-fault</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>Triple-fault</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

#### 3.2 Synthetic Track

To present the scenario generation algorithm with the appropriate level of formality we need a number of definitions.

**Definition 1.** (Diagnostic System). A diagnostic system DS is defined as the triple DS = ⟨SD, COMP, OBS⟩, where SD is a propositional theory over a set of variables V, COMP ⊆ V, OBS ⊆ V, COMP is the set of assumables, and OBS is the set of observables.

We partition the set of observable variables OBS into inputs IN and outputs OUT such that OBS = IN ∪ OUT and IN ∩ OUT = ∅.

**Definition 2.** (Diagnosis). Given a diagnostic system DS = ⟨SD, COMP, OBS⟩, an observation α over some variables in OBS, and a health assignment ω, ω is a diagnosis iff SD ∧ α ∧ ω is consistent.

**Definition 3.** (Minimal Diagnosis). A diagnosis ω is minimal if no diagnosis ω’ exists such that NL(ω’) ⊂ NL(ω), where NL(ψ) is the set of negative literals in ψ.

**Definition 4.** (Cardinality of a Diagnosis). The cardinality of a diagnosis, denoted as |ω|, is defined as the number of negative literals in ω.

A minimal cardinality diagnosis is a minimal diagnosis, but the opposite does not hold. There are minimal diagnoses which are not minimal cardinality diagnoses.

The purpose of Alg. 1 is to generage observations leading to diagnoses of increasing minimal cardinality.

#### Algorithm 1: A greedy stochastic scenario generation algorithm

**Function** MAKEALPHAS(DS, N ) returns set of terms

**Inputs:**
- DS = ⟨SD, COMP, OBS⟩, diag. system
- OBS = IN ∪ OUT, IN ∩ OUT = ∅
- N, integer, observations per cardinality

**Local variables:**
- α, β, α̲, fault, terms
- i, c, integers,
- R, set of terms, result, initially ∅

1: for i = 1 ... N do
2: α ← RANDOMINPUTS(IN)
3: β ← COMPUTENOMINALOUTPUTS(DS, α)
4: c ← 0
A set of 9 metrics has been defined for assessing the performance of the diagnostic algorithms. For DXC we make a distinction between temporal, technical, and computational performance metrics. The temporal metrics measure how quickly an algorithm responds to faults in a physical system. The technical metrics measure non-temporal features of a diagnostic algorithm including accuracy and diagnostic cost/utility. Finally, computational metrics are intended to measure how efficiently an algorithm uses the available computational resources.

In addition, we divide the metrics into 2 main categories:

Detection metrics which deal with temporal, technical, and computational metrics associated with only detection of the fault.

Isolation metrics which deal with temporal, technical, and computational metrics associated with isolation of the fault.

The 9 metrics are listed in Table 6. The notation used for the definition of the metrics is as follows:

S – The set of scenarios for a given system description
Sn – The set of nominal scenarios for a given system description
Sf – The set of faulty scenarios for a given system description
tfd – The time when the fault detection signal has been asserted for the first time
ti – The time when the last persistent fault isolation signal has been asserted
ωact – The true component mode vector (ground truth)
\( \omega_{\text{prec}} \) – The predicted component mode vector (represents the set of candidate diagnoses by the DA)

\( T_d \) – Total computation time

\( M_d \) – Peak amount of allocated memory

\( C \) – All possibly faulted components

\( D \) – Faulted components in \( \omega_{\text{prec}} \)

\( I \) – Faulted components in \( \omega_{\text{act}} \)

Finally, using the aforementioned notation, the 9 metrics are defined as:

**M_{fd} – Fault Detection Time:** The reaction time for a diagnostic engine in detecting an anomaly (Kurtoglu et al., 2008).

\[ M_{fd} = t_{fd} \]  

**M_{fi} – Fault Isolation Time:** The time for isolating a fault (Kurtoglu et al., 2008). In many applications this metric is less important than the diagnostic accuracy, but it is important in sequential diagnosis, probing, etc.

\[ M_{fi} = t_{fi} \]  

**M_{fp} – False Positive Rate:** The metric that penalizes diagnostic algorithms which announce spurious faults (Kurtoglu et al., 2008). The false positive rate is defined as:

\[ M_{fp} = \frac{\sum_{s} m_{fp}(s)}{|S|} \]  

where for each scenario \( s \) the “false positive” function \( m_{fp}(s) \) is defined as:

\[ m_{fp}(s) = \begin{cases} 1, & \text{if } t_{da} < t_{s} \\ 0, & \text{otherwise} \end{cases} \]  

where \( t_{s} = \infty \) for a nominal scenario

**M_{fn} – False Negative Rate:** The metric that measures the ratio of missed faults by a diagnostic algorithm (Kurtoglu et al., 2008).

\[ M_{fn} = \frac{\sum_{s \in S} m_{fn}(s)}{|S|} \]  

where for each scenario \( s \) the “false negative” function \( m_{fn}(s) \) is defined as:

\[ m_{fn}(s) = \begin{cases} 1, & \text{if } t_{s} = \infty \\ 0, & \text{otherwise} \end{cases} \]  

**M_{fa} – Detection Accuracy:** The fault detection accuracy is the ratio of the number of correctly classified cases to the total number of cases (Kurtoglu et al., 2008). It is defined as:

\[ M_{fa} = 1 - \frac{\sum_{s \in S} m_{fp}(s) + m_{fn}(s)}{|S|} \]  

**M_{ce} – Classification Errors:** Isolation classification error metric measures the accuracy of the fault isolation by a diagnostic algorithm and is defined as the Hamming distance between the true component mode vector \( \omega_{\text{act}} \) and the predicted component mode vector \( \omega_{\text{prec}} \).

In the calculation of the classification error metric, the data values for the Hamming distance are the respective modes of components comprising a system description. For example, if the true component mode vector of the system is \([1,0,0,1,0]\) and the predicted component mode vector is \([1,1,0,0,0]\), the classification error is 2. If more than one predicted mode vector is reported by a DA, (meaning that the diagnostic output consists of a set of candidate diagnoses), then the classification error is calculated for each predicted component mode vector and weighted by candidate probabilities reported by the DA.

**M_{au} – Diagnostic Utility:** The intuition behind the metric is to charge a DA for every incorrect component replacement it required to restore the circuit to functioning. For example, the correct diagnosis should always receive a perfect score. The diagnosis all components bad has a cost of the number of components. Consider a single fault and the DA reports all components good. Finding the faulty component would require on average replacing component by component until the system was functioning correctly (on average half the components). More generally:

\[ M_{au} = l|I| / (|I| + c (|I| - D, |C - D|)) \]  

Where \( c(n, m) \) is defined as the expected number of trials needed to isolate \( n \) out of \( m \). If \( n \) is much smaller than \( m \), then it is approximately:

\[ c(n, m) = nm / (n + 1) \]  

For example, to find 1 fault in \( m \) has cost \( m/2 \). To find 2 faults in \( m \) is \( 2m/3 \). Similarly to the classification metric, if more than one predicted mode vector is reported by a DA, then error is calculated for each predicted component mode vector and weighted by candidate probabilities reported by the DA.

**M_{cpu} – CPU Load:** This is the average CPU load during the experiment

\[ M_{cpu} = t_{s} + \sum_{q \in S} q \]  

where \( t_{s} \) is the startup time of the diagnostic engine and \( T_d \) is a vector with the actual CPU time spent by the diagnostic algorithm at every time step in the diagnostic session.

**M_{mem} – Memory Load:** This is the maximum memory size at every step in the diagnostic session. CPU load during the experiment

\[ M_{mem} = \max_{m \in M_d} m \]  

where \( M_d \) is a vector with the maximum memory size at every step in the diagnostic session.

---

\( ^3 \) The Hamming distance between two strings of data values (of equal length) is the number of positions for which the corresponding data values are different.
5. COMPETITION SETUP AND SCORING

Version 1.1 of the DXC Framework, implemented as specified in (Kurtoglu et al., 2009), was used to run the competition. Two computers with identical hardware\textsuperscript{4} were set up, one running Windows\textsuperscript{\textregistered} and the other Linux. The choice of target operating system was left to DA developers.

System profiling was performed on the machines over a period of days to ensure stable experiment conditions.

DAs were run on competition datasets over a period of two weeks. The Evaluator was then run on the full results set, assigning relative rankings for each metric. Since there were multiple systems in the Synthetic Track, the metrics computed for each system were aggregated before assigning relative rankings. The per scenario metrics were averaged over all scenarios and aggregated over all systems. For each of the Industrial track tiers there was only one system, so no aggregation was necessary.

A DA that ranked first place in a given metric was awarded 10 points, second place was awarded 8, third 7, etc. This score was then multiplied by a metric weight, shown in Tables 8, 9, and 10, and added to the DA’s total.

Metric weights for the Industrial Track were determined by considering a number of use cases in which the importance of each metric was subjectively assessed. For example, in an abort use case high importance was given to the mean time to detect a fault whereas in a maintenance use case more weight was given to the ability to isolate a fault. Similar considerations were given to use cases such as real-time recovery and control, ground support operations, and resource limited applications. Since a use case was not specified as part of the competition scenarios, we simply averaged over all of the use cases to arrive at the final metric weights.

6. DIAGNOSTIC ALGORITHMS

The teams that participated in the First International Diagnosis Competition are listed in Table 7.

<table>
<thead>
<tr>
<th>Team Name</th>
<th>Track(s)</th>
<th>Algorithm Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>FACT</td>
<td>I1, I2</td>
<td>Model-based</td>
</tr>
<tr>
<td>Fault Buster</td>
<td>I1, I2</td>
<td>Statistical</td>
</tr>
<tr>
<td>HyDE-A</td>
<td>I1, I2</td>
<td>Model-based</td>
</tr>
<tr>
<td>HyDE-S</td>
<td>I1</td>
<td>Model-based</td>
</tr>
<tr>
<td>Lydia</td>
<td>S</td>
<td>Model-based</td>
</tr>
<tr>
<td>NGDE</td>
<td>S</td>
<td>Model-based</td>
</tr>
<tr>
<td>ProADAPT</td>
<td>I1, I2</td>
<td>Probabilistic</td>
</tr>
<tr>
<td>RacerX</td>
<td>I1</td>
<td>Change detection</td>
</tr>
<tr>
<td>RODON</td>
<td>I1, I2, S</td>
<td>Model-based</td>
</tr>
<tr>
<td>RulesRule</td>
<td>I1</td>
<td>Rule-based</td>
</tr>
<tr>
<td>StanfordDA</td>
<td>I2</td>
<td>Optimization</td>
</tr>
</tbody>
</table>

\textsuperscript{4} Intel\textsuperscript{\textregistered} XEON\textsuperscript{\textregistered} 2x2.20Ghz, 3.60 GB RAM

A total of twelve DAs participated, nine in Tier 1 of the Industrial Track, six in Tier 2, and three in the Synthetic Track. Brief descriptions of each of these algorithms are provided below:

1. FACT – a model-based diagnosis system that uses hybrid bond graphs, and models derived from them, at all levels of diagnosis, including fault detection, isolation, and identification. Faults are detected using an observer-based approach with statistical techniques for robust detection. Faults are isolated by matching qualitative deviations caused by fault transients to those predicted by the model. For systems with few operating configurations, fault isolation is implemented in a compiled form to improve performance (Roychoudhury et al., 2009).

2. Fault Buster – is based on a combination of multivariate statistical methods, for the generation of residuals. Once the detection has been done a neural network performs classification for doing isolation.

3. HyDE-A – HyDE (Hybrid Diagnosis Engine) is a model-based diagnosis engine that uses consistency between model predictions and observations to generate conflicts which in turn drive the search for new fault candidates. HyDE-A uses discrete models of the system and a discretization of the sensor observations for diagnosis (Narasimhan and Brownston, 2007).

4. HyDE-S – uses the HyDE system but runs it on interval values hybrid models and the raw sensor data (Narasimhan and Brownston, 2007).

5. Lydia – is a declarative modeling language specifically developed for Model-Based Diagnosis (MBD). The language core is propositional logic, enhanced with a number of syntactic extensions for ease of modeling. The accompanying toolset currently comprises a number of diagnostic engines and a simulator tool (Feldman et al., 2006).

6. NGDE – Allegro Common Lisp implementation of the classic GDE. Uses a minimum-cardinality candidate generator to construct diagnoses for the competition.

7. ProADAPT – processes all incoming environment data (observations from a system being diagnosed), and acts as a gateway to a probabilistic inference engine. It uses the Arithmetic Circuit (AC) Evaluator which is compiled from Bayesian network models. The primary advantage to using ACs is speed, which is key in resource bounded environments (Mengshoel 2007).

8. RacerX – is a detection-only algorithm which detects a percentage change in individual filtered sensor values to raise a fault detection flag.

9. RODON – is based on the principles of the General Diagnostic Engine (GDE) as described by de Kleer and Williams and the G+DE by Heller and Struss. RODON uses contradictions (conflicts) between the simulated and the observed behavior to generate hypotheses about possible causes for the observed behavior. If the model
10. RulesRule – is a rule-based isolation-only algorithm. The rule base was developed by analyzing the sample data and determining characteristic features of fault. There is no explicit fault detection though isolation implicitly means that a fault has been detected.

11. StanfordDA – is an optimization-based approach to estimating fault states in a DC power system. The model includes faults changing the circuit topology along with sensor faults. The approach can be considered as a relaxation of the mixed estimation problem. We develop a linear model of the circuit and pose a convex problem for estimating the faults and other hidden states. A sparse fault vector solution is computed by using $l_1$ regularization (Zymnis et al., 2009).

12. Wizards of Oz – is a consistency-based algorithm. The system of the model completely defines the stable (static) output of the system in case of normal and faulty behavior. Given a new command or new observations, the algorithm waits for a stable state and computes the minimum diagnoses consistent with the observations and the previous diagnoses.

7. RESULTS AND DISCUSSION

7.1 Industrial Track

The results for the Industrial Track are shown in Table 8 and Table 9 for Tier 1 and Tier 2, respectively. The overall winner for both tracks was ProADAPT. RODON placed second in Tier 1 and third in Tier 2. The StanfordDA, which did not participate in Tier 1, placed second in Tier 2. However, ProADAPT and StanfordDA benefitted from previous funded experience with ADAPT so RODON was the official winner of both tiers. The distribution of first or second ranks within each metric was spread out among the DAs, no DA ranked first or second for all of the metrics. Note that the final scores and ranks depend on the weights applied to each metric. Different weights, corresponding to different use cases, would affect the results. The sensitivity of the results to the metrics and weights is left for future study.

Figures 2-9 are graphical depictions of the data in Tables 8 and 9. A few observation follow. False positives were counted in the following two situations: for nominal scenarios where the DA declared a fault; and for faulty scenarios where the DA declared a fault before any fault was injected. An error in the rule base of RulesRule led to more false positive indications for the faulty scenarios than for the nominal scenarios and also resulted in a large number of classification errors. For other DAs, false positives also resulted from nominal commanded mode changes in Tier 2 in which the relay feedback did not change status as of the next data sample after the command. Here is an extract from one of the input scenario files that illustrates this situation:

```plaintext
command @120950 EY275_CL = false;
sensors @121001 {... ESH275 = true, ...}
sensors @121501 {... ESH275 = false, ...}
```

A command is given at 120.95 seconds to open relay EY275. The associated relay position sensor does not indicate open as of the next sensor data update 51 milliseconds later. This is nominal behavior for the system and examples were provided in the sample data. A DA that does not account for this delay will likely indicate a false positive in this case.

In several instances DAs reported diagnosis mode IDs which contains failure modes besides the nominal behavior, these can be used to verify the hypotheses, which speed up the diagnostic process and improve the results (Karin et al., 2006).
did not match the names specified in the system catalog. For these cases the diagnosis was treated as an empty candidate. This could either negatively or positively impact the classification error metric depending on whether the DA had a correct or incorrect isolation. Participants were encouraged to run their DA output through the evaluator code that was distributed with the sample data sets to check for and correct these syntax errors.

There are a few remarks in regards to the timing metrics listed in Table 6 and shown graphically in Fig. 4. First, RacerX did not have an isolation time as it was a detection-only DA. Second, note the somewhat confusing result that the mean isolation time for RulesRule was less than the mean detection time. This has to do with the way the metrics are calculated. The detection time is undefined for scenarios with a false positive; however, the isolation time is not necessarily undefined and is calculated as discussed in section 4. The intent is to account for the situation where a DA retracts a spurious detection signal and subsequently isolates to the correct component. In this case the scenario is declared a false positive but the accuracy and timing of the isolation is calculated with respect to the last persistent diagnosis. Consequently, for DAs with many false positives the detection time may be calculated for fewer scenarios than the isolation time with the result that the mean isolation time for all scenarios could be less than the mean detection time. However, in any scenario where both times are defined, the DA isolation time is always greater than or equal to the detection time, as would be expected.

Tier 1 had the interesting circumstance that the same DA was implemented by two different modelers. HyDE-A was modeled primarily with Tier 2 in mind and had a policy of waiting for transients to settle before requesting a diagnosis. The same policy was simply applied to Tier 1 as well, even though transients in Tier 1 corresponded strictly to fault events. On the other hand, HyDE-S was modeled only for Tier 1 and did not include a lengthy time-out period for transients to settle. HyDE-S had dramatically smaller mean detection and isolation times (see Fig. 4) with roughly the same number of classification errors (Fig. 3) as HyDE-A. This illustrates the kind of impact that modeling and implementation decisions have on DA performance.

7.2 Synthetic Track

As can be seen in Table 7 all synthetic track DAs are model-based. Lydia uses a stochastic approach to identify diagnoses while RODON and NGDE use the familiar GDE-like approaches. Their overall utility scores are not dramatically different.

The results for the Synthetic Track are presented in Table 10. Based on the overall metric NGDE was first, Lydia second, and RODON third. Lydia was used to generate the scenario sets and therefore is disqualified. Furthermore the designers of Lydia and NGDE both participated in the design of DXC, and are thus disqualified. So RODON is the official winner. RODON scored reasonably well on the smaller circuits but failed to return any diagnoses for the 4 larger circuits.

Use of computational resources varied dramatically over the systems. Lydia used an order of magnitude fewer resources than either RODON or NGDE and thus ranked first along the memory and CPU metrics. RODON and NGDE are very similar in resource usage, with RODON edging out NGDE.

Fig. 10 shows the DA utility for each of the circuits. Note that the utility score decreases significantly with circuit size.

<table>
<thead>
<tr>
<th>FP Rate</th>
<th>Weight</th>
<th>RODON</th>
<th>Wizards Of Oz</th>
<th>Fault Buster</th>
<th>ProADAPT</th>
<th>HyDE</th>
<th>Stanford</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking Points</td>
<td>1.3</td>
<td>0.5417</td>
<td>0.5106</td>
<td>0.8143</td>
<td>0.0732</td>
<td>0.0000</td>
<td>0.3256</td>
</tr>
<tr>
<td>FN Rate</td>
<td>1.3</td>
<td>0.0972</td>
<td>0.0959</td>
<td>0.2400</td>
<td>0.1392</td>
<td>0.3000</td>
<td>0.0519</td>
</tr>
<tr>
<td>T_det (ms)</td>
<td>2.2</td>
<td>3490</td>
<td>30742</td>
<td>14099</td>
<td>5981</td>
<td>17610</td>
<td>3946</td>
</tr>
<tr>
<td>T_iso (ms)</td>
<td>1.5</td>
<td>36331</td>
<td>47625</td>
<td>37808</td>
<td>12486</td>
<td>21982</td>
<td>14103</td>
</tr>
<tr>
<td>CPU (ms)</td>
<td>0.6</td>
<td>80251</td>
<td>23387</td>
<td>5798</td>
<td>3416</td>
<td>29612</td>
<td>963</td>
</tr>
<tr>
<td>Mem (kb)</td>
<td>0.6</td>
<td>29878</td>
<td>7498</td>
<td>10261</td>
<td>6539</td>
<td>20515</td>
<td>5912</td>
</tr>
<tr>
<td>FINAL RANK:</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
This decrease is not a result of poor performance or algorithm design. Rather, an oracle could not do much better as a large number of faults can exhibit the same input-output behavior and no DA could isolate the injected fault out of the large ambiguity groups. The challenge presented by large ambiguity groups is discussed further in the following section and the NGDE and Lydia papers included in this collection.

8. ASSUMPTIONS, ISSUES, AND EXTENSIONS

The primary goal of this competition was to demonstrate an end-to-end implementation of the DXC framework and create a foundation for future DX competitions. As a result we made several simplifying assumptions. We also ran into several issues during the course of this implementation that could not be addressed. In this section, we try to present those assumptions and issues, which we hope can be addressed in future competitions.

Although the competition was a success, it only addresses a small set of the types of diagnostic tasks, which occur in practice. It would be unfortunate for the DX community to focus only on the tasks of this competition. Our goal is to continually expand the coverage of diagnostic challenges experienced in the field. Our hope is that every successive year will expand the set of tasks in the competition and in doing so produce an ever growing repository DX researchers have available to evaluate their own algorithms.

8.1 Competition Scope

In the first year of the diagnostic competition, the fault signatures were limited to abrupt parametric and discrete types. Faults were inserted assuming uniform probabilities and included component and sensor faults. In future years, we will provide the failure rates of components and use these to evaluate the precision of diagnoses. For the Industrial Track, other fault types are presently possible to inject in the testbed – including incipient, intermittent, and noise – and could be included in future work. Additional ideas for future research include giving DAs reduced sensor sets, introducing multirate sensor data, injecting transient faults, allowing for autonomous transitions, adding variable loads, and extending the scope and complexity of the physical system. For the synthetic track, all the systems were known a priori. This means researchers could optimize for these circuits. We don’t believe this happened this year, but to avoid this in future years we will include entirely novel circuits along with the familiar ones. This year we sampled only one observation time. We will provide multiple observations. This will evaluate a DAs ability to merge information from multiple times. An important component of troubleshooting is introducing probe points. In future years, we can evaluate the number of probes needed to isolate the fault. This year the input vector was supplied. The diagnostician could construct the input vector, which was most informative. This year the Synthetic Track focused on combinatorial circuits. In subsequent years we hope to introduce troubleshooting of sequential circuits. Finally, digital circuits are convenient to model and conveniently illustrate many aspects of diagnostic algorithms. In future years, we will extend the types of

<table>
<thead>
<tr>
<th>circuit</th>
<th>#comp</th>
<th>Lydia</th>
<th>NGDE</th>
<th>RODON</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>cpu</td>
<td>mem</td>
<td>utl</td>
</tr>
<tr>
<td>74182</td>
<td>19</td>
<td>0.4137</td>
<td>6335</td>
<td>0.4793</td>
</tr>
<tr>
<td>74L85</td>
<td>33</td>
<td>0.2433</td>
<td>6365</td>
<td>0.3098</td>
</tr>
<tr>
<td>74283</td>
<td>36</td>
<td>0.1580</td>
<td>6385</td>
<td>0.1553</td>
</tr>
<tr>
<td>74181</td>
<td>65</td>
<td>0.1504</td>
<td>6619</td>
<td>0.1931</td>
</tr>
<tr>
<td>c432</td>
<td>160</td>
<td>0.0871</td>
<td>7520</td>
<td>0.2096</td>
</tr>
<tr>
<td>c499</td>
<td>202</td>
<td>0.0622</td>
<td>20347</td>
<td>0.0699</td>
</tr>
<tr>
<td>c880</td>
<td>383</td>
<td>0.0483</td>
<td>13718</td>
<td>0.0401</td>
</tr>
<tr>
<td>c1355</td>
<td>546</td>
<td>0.0295</td>
<td>22550</td>
<td>0.0246</td>
</tr>
<tr>
<td>c1908</td>
<td>880</td>
<td>0.0179</td>
<td>26171</td>
<td>0.0150</td>
</tr>
<tr>
<td>c2670</td>
<td>1193</td>
<td>0.0647</td>
<td>20537</td>
<td>0.1076</td>
</tr>
<tr>
<td>c3540</td>
<td>1669</td>
<td>0.0319</td>
<td>27022</td>
<td>0.0407</td>
</tr>
<tr>
<td>c5315</td>
<td>2307</td>
<td>0.0165</td>
<td>30926</td>
<td>0.0275</td>
</tr>
<tr>
<td>c6288</td>
<td>2416</td>
<td>0.0008</td>
<td>17483</td>
<td>0.0563</td>
</tr>
<tr>
<td>c7552</td>
<td>3512</td>
<td>0.0317</td>
<td>37989</td>
<td>0.0283</td>
</tr>
<tr>
<td>Averaged</td>
<td>1613</td>
<td>0.0969</td>
<td>17855</td>
<td>0.1255</td>
</tr>
<tr>
<td>Per Metric Rank</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Points</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Metric Weight</td>
<td>1.5</td>
<td>1.5</td>
<td>7</td>
<td>1.5</td>
</tr>
<tr>
<td>Final Scores</td>
<td>86</td>
<td>91</td>
<td></td>
<td>73</td>
</tr>
<tr>
<td>Final Rank</td>
<td>2</td>
<td>1</td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>
systems to include. Two comparatively easy types of systems
to add are reprographic engines as we have a tool available to
generate such models, and analog circuits.

8.2 Metrics

Selecting the set of metrics to be used for evaluation was a
challenging job. We based our decision on the system and
kinds of faults we were dealing with. In reality we also need
to design metrics more closely associated with the context of
use. One common metric is to minimize total cost of repair
where cost includes down time to the customer, diagnostician’s
time, parts, etc. In addition since we were
dealing with abrupt, persistent, and discrete faults, metrics
associated with incipient, intermittent, and/or continuous
faults were not considered. The metrics listed in this paper do
not capture the amount of effort necessary to build models of
sufficient fidelity for the diagnosis task at hand. Furthermore,
we did not attempt to investigate the ease or difficulty of
updating models with new or changed system information.
The art of building models is an important practical
consideration which is not addressed in the current work.

The isolation accuracy metric used for the industrial track was
not suitable for the synthetic track. A DA which reported
nothing wrong on every scenario would come close to
winning the competition based on this metric. The main
problem with this metric is that the number of faulty
components is always small with respect to the size of the
system. As a result we cannot differentiate adequately
between a few faults and no faults.

Isolation classification error was also not suitable for the
synthetic track. This metric still suffers from the problem that
all good is scored too high: The Hamming distance between a
single fault and every component good is very small!

Ideally we would like to use a SAT solver to evaluate the
accuracy of a DA’s diagnosis. However, we did not have time
to implement it so as an alternate we selected utility as the
isolation accuracy metric for the synthetic track. One of the
major flaws of this metric is that average expected utility
scores decrease with system size, thereby implicitly de-
weighting diagnoses of larger circuits. We also considered and
rejected a classification error metric which would assign a
high score to any fault from an ambiguity group which we
considered a bigger flaw than the flaw for the utility metric.

Finally, the current isolation metrics evaluate diagnostic
performance based on a discrete isolation assumption in
which faults are isolated to one of the discrete modes of a
component. As more continuous type faults are introduced,
additional or generalized metrics are required in order to
calculate the accuracy of isolation estimates on a continuous
scale.

8.3 Competition Setup

Some practical issues arose in the execution of competition
experiments. Much effort was put into ensuring stable,
uniform conditions on the host machines; however, due to
time constraints and the unpredictable element introduced by
running external DA submissions, it was necessary to take
measures that may have caused slight variability. One
example was the manual examination of ongoing experiment
results for quality assurance. Future releases of the DXC
Framework can address this by being more robust to
unexpected DA behavior, and sending email notifications in
the event of such.

Additionally, for Java DAs, significant differences were
evident in the peak memory usage metric when run on Linux
versus Windows™. The cause for this was not explored due to
time constraints, as the method used on Windows™ for
calculating peak memory usage involved a Windows™ API
system call, the analysis of which was deemed too expensive.

The problem was bypassed by running all Java DAs on Linux.
This worked for all save one, RODON. When it was
determined that any change in RODON’s peak memory usage
score would not affect the final rankings in any way, the issue
was waived.

9. CONCLUSIONS

We presented the successful implementation of the DXC
framework called DXC’09. We learned some valuable lessons
trying to run this competition. One major takeaway is that
there is still a lot of work and discussion needed to determine
common comparison and evaluation framework for the
diagnosis community.

We hope to continue the work next year by running DXC’10.
We have identified several ways to extend the systems used in
the current competition some of which can be achieved in a
year’s time. We also hope to add other systems to the fold,
which may pose different diagnostic challenges.

ACKNOWLEDGMENTS

We extend our gratitude to David Nishikawa (NASA), David
Jensen (Oregon State University), Brian Ricks (University of
Texas at Dallas), Ole Mengshoel (Carnegie Mellon
University), Adam Sweet (NASA), David Hall (Stinger
ghaffarian Technologies), all DXC’09 competitors, the
DX’09 organizers and many others for valuable discussions,
criticism and help.

REFERENCES

Bartys, M., R. Patton, M. Syfert, S. de las Heras, and J.
Quevedo (2006). Introduction to the DAMADICS
Actuator FDI Benchmark Study, Control
Engineering Practice, vol 14, pp. 577-596.
combinational benchmark circuits and a target translator
de Kleer, J (2008). An improved approach for generating
DX’08, pp. 247–252.
roads lead to fault diagnosis: Model-based reasoning
with LYDIA. In: Proc. BNAIC’06.
Computing observation vectors for Max-Fault Min-
924.


Appendix A. FIGURES.

Fig. 1. The ADAPT EPS (Electrical Power System)
Fig. 2. Industrial track tier 1 false positive rate, false negative rate, and detection accuracy by DA

Fig. 3. Industrial track tier 1 classification errors by DA

Fig. 4. Industrial track tier 1 detection and isolation times by DA

Fig. 5. Industrial track tier 1 CPU time and peak memory usage by DA

Fig. 6. Industrial track tier 2 false positive rate, false negative rate, and detection accuracy by DA

Fig. 7. Industrial track tier 2 classification errors by DA
Fig. 8. Industrial track tier 2 detection and isolation times by DA

Fig. 9. Industrial track tier 2 CPU time and peak memory usage by DA

Fig. 10. Synthetic track DA utility scores by circuit