SAND2017-3099C

OPENFABRICS ALLIANCE

13th ANNUAL WORKSHOP 2017

HOST BASED INFINIBAND NETWORK FABRIC MONITORING

Michael Aguilar

Jim Brandt Staff R&D, Ann Gentile, Benjamin Allan, and Douglas Pase

Sandia National Laboratories

[March, 2017]

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2013-6521C

OUTLINE

- **Motivation**
  - Infiniband network fabric problems, including contention, congestion, and their sources are not well understood
  - Current InfiniBand fabric monitoring tools and techniques have too low fidelity and lack sufficient synchronization to be of practical use in understanding congestion phenomena
- **Goals**
  - Understand network data flow as it is routed through IB switches.
  - Timely and synchronized identification of network congestion within an IB fabric.
    - Better understanding of factors that cause congestion
    - Develop tools and techniques for early identification and mitigation of network contention related problems
- **Related Work**
  - Need to cite other work here – e.g. subnet manager,
  - distributed script using ibstat
  - LLNL approach (can't remember the name), etc.
  - LDMS host metrics collected from local host HCAs using Sysclassib sampler
  - Each metric sample can be gathered synchronously at sub-second intervals
- **Approach**
  - Create a distributed LDMS sampler to "synchronously" retrieve metrics from all of switch ports within cluster Infiniband fabric.
    - Synchronous to some delta determined by a switch's port count and processor capabilities
    - The sampler provides network connection information, as well
- **Initial Results**
  - How long does it take to get metrics from switch ports and what is the load on the system?
  - Metrics retrieved by sampler.
- **Upgrades and Future Work**
- **Conclusion**

OUTLINE

- **Motivation**
  - Understand network data flow as it is routed through IB switches.
  - Timely and synchronized identification of network congestion within an IB fabric.
    - Better understanding of factors that might cause congestion.
- **Related Work**
  - LDMS metrics gained at data source and drain through the use of Sysclassib sampler
    - Each Sysclassib sampler gathers metrics from locally connected compute node HCAs
  - Each metric sample can be gathered synchronously at shorter than 1 second intervals.
- **Approach**
  - A new sampler is being created to retrieve metrics from all of the switch ports within cluster Infiniband fabrics.
    - Information can be gathered synchronously to provide metrics on data flowing throughout the network, from Source to Drain.
  - The sampler provides network connection information, as well
- **Initial Results**
  - How long does it take to get metrics from switch ports and what is the load on the system?
  - Metrics retrieved by sampler.
- **Upgrades and Future Work**
- **Conclusion**

MOTIVATION

MOTIVATION

- **Help improve computational performance by better network routing, data locality, and reduction of congestion.**
- **Understand network data flow as it is routed through Infiniband Switches**
  - HPC systems depend upon good quality internal RDMA data flow for computational performance.
    - Internal RDMA data transfers, using Infiniband, often are made through several layers of switches and gateways on very large HPC systems.
    - Application data-flows are often dynamic in volume from source to drain.
      - Counter monitoring at the RDMA source and drain only provide data that shows activity at the endpoints. This leaves details of routing information and congestion, that might occur internally, out of monitoring data.
- **Timely and synchronized identification of network congestion within an IB fabric.**
  - Better understanding of factors that might cause congestion.
    - Guesses must be made of what is happening in switches and gateways because traffic flows are hidden from view.
      - “Educated” guesses must be used to create policies and actual wiring.
      - Over-design to provide wider data highways must be done to counteract congestion created during RDMA exchanges.

INFINIBAND AND MAD QUERIES CAN BE PERFORMED FROM ANY HCA WITHIN THE CONNECTED FABRIC

- **Metrics can be read using MAD queries through VL15 on any LID and port within the fabric.**
  - A sampler can be created to read RDMA network transmission metrics within the fabric
    - By gathering full data metrics from switches within the fabric, a more complete picture can be gathered of how much data is traversing each switch port
- **The load of monitoring the HCAs and connections in the different layers of switches can be spread across all nodes.**
  - The connectivity is itself important metric information.
  - OFED libnetdisc can be use to find connectivity and determine how to divide up the query tasks.

RELATED WORK

RELATED WORK

- **The LDMS sampler Sysclassib provides timely data-gathering at an RDMA Source and Drain**
  - At each compute node, the Sysclassib sampler reads metrics from the local HCA LID and port.
    - Data gathered includes packets, bytes, 32-bit congestion metrics, and wait data
    - Each sample can be gathered at intervals of less than 1 second.
      - Fine-grained rates of change of data flows can be seen on each compute node that show how much data is transiting from tightly-coupled application threads.
        - » Can give an idea of how much of a resource is being used during a scheduled batch run of an application computation.
  - With better information on data transmission activity at the Source and Drain HCAs, MPI migrations and source-initiated delay injection can be done to help alleviate congestion.
    - On the LANL Cielo system, fine-grained monitoring of end network connections has allowed running applications to migrate from compute node to compute node to reduce network congestion.
- **INAM monitoring allowed monitoring of the entire subnet, including switches.**
  - However only 1 port at a time could be analyzed by the monitoring system.
    - Dynamic gathering of data metrics could not be done for running applications.

APPROACH

- **The objective of the new monitoring system was to report back a detailed set of metrics from the entire Infiniband fabric during a sample cycle.**
  - We were able to gather metrics from the Source and Drain, so but we wanted to gather metrics from the switch ports, as well
    - By sampling data from switches within an HPC system and all active HCAs at the end-points, detailed information on data flows can be analyzed.
      - The user can specify the subset of metrics to collect. If a set of metrics isn't chosen, the default setting is to display all of the metrics.
      - To allow easy parsing of resulting data, all of the metrics should display the LID and port identifiers.
      - Allow the user to pick more local switch ports to check from the sampler HCA.
- **Sampling should be time-synchronized .**
  - LDMS samplers maintain a time-synchronized metric gathering and report when there are synchronization issues.
  - With good time synchronization, rate changes in metric values can be ascertained.

DELEGATOR DAEMON

- **A delegator daemon gathers information on network connections and available LIDs and ports.**
  - With the LIDs and ports gathered, each sampler is then provided with a subset of the network to gather metrics from.
    - Division of labor within the subnet.
    - Each sampler was provided a list of LIDs and ports, the brand and type of network port, and the connecting LID and port
  - Samplers were given their identities and work to be performed by their component identifier numbers.
    - An input file was used to list the metrics that the LDMS sampler was to gather from each port.
    - An input parameter would provide the delegator with information on the brand and type of Infiniband hardware to expect.
    - Finally, an input file containing both the quantity of samplers used in metric gathering and the some user-defined LIDS and ports to sample from.
      - Allowed the LDMS IBFabric sampler to gather metrics from locally connected switches.

DELEGATOR DAEMON

- **Sample Metric Selection File with # characters as comments**

```
Remote_LID
Remote_port
#symbol_error
#link_error_recovery
#link_downed
#port_rcv_errors
#port_rcv_remote_physical_errors
#port_rcv_switch_relay_errors
#port_xmit_discards
#port_xmit_constraint_errors
#port_rcv_constraint_errors
#COUNTER_SELECT2_F
#local_link_integrity_errors
#excessive_buffer_overrun_errors
#VL15_dropped
port_xmit_data
port_rcv_data
port_xmit_packets
port_rcv_packets
#port_xmit_wait
#port_unicast_xmit_packets
#port_unicast_rcv_packets
#port_multicast_xmit_packets
#port_multicast_rcv_packets
```

- **Sample Configuration file with LID assignment to directly connected LIDs and Ports**

```
samplers  16
14 14:3 21:25 21:26 21:27 21:28 14:25
```

SAMPLERS

- During configuration, the new IBFabric sampler communicates its component identifier to the delegator then gets a group of LIDs and ports to check.
- Using the LID and port numbers, MAD queries were performed using LID and port identifiers and the data was gathered into a complete LDMS sampling set.
  - Each time the sampler gathers metrics from a switch port, a full set of information is retrieved using a MAD query
  - The MAD query results are decoded using more library calls.
  - Depending upon the type of results that the LDMS user requests, an LDMS metric set is then compiled to later retrieval by an LDMS Aggregator
    - The LDMS set size is designed to be dynamically allocated by the number of LIDs and ports to sample and the type of samples to be done.

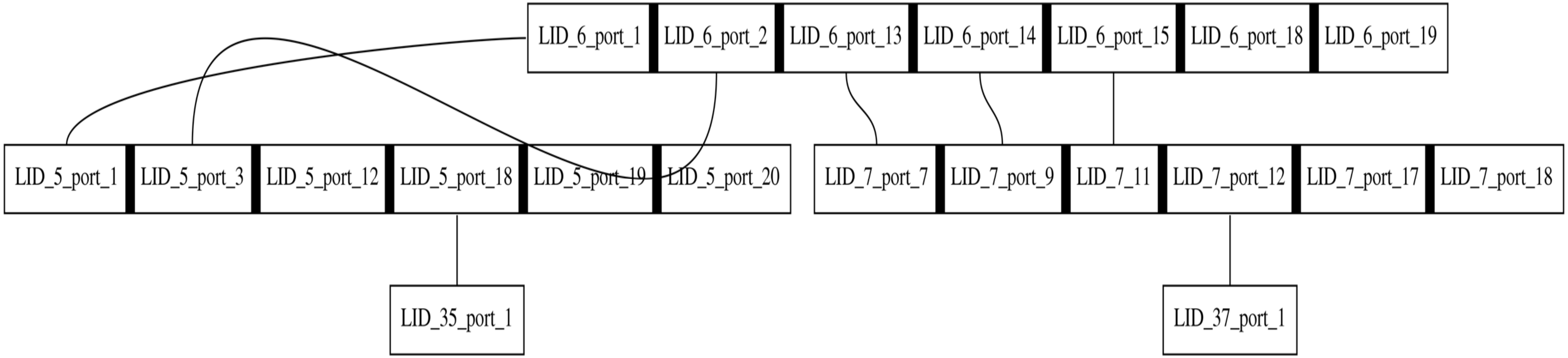
SAMPLERS

- **Sample LDMS metric output from IBFabric using the sample input files from above:**

```
[ METADATA --------
    Producer Name : shaun3
    Instance Name : shaun3/ibfabric
      Schema Name : ibfabric_shaun3
             Size : 1600
     Metric Count : 26
               GN : 2
  DATA ------------
        Timestamp : Thu Mar 16 14:19:56 2017 [3159us]
         Duration : [0.000533s]
       Consistent : TRUE
             Size : 248
               GN : 150
  -----------------
M u64      component_id                    3
D u64      LID_21_port_28_Remote_LID        14
D u64      LID_21_port_28_Remote_port       28
D u64      LID_21_port_28_port_xmit_data    17424
D u64      LID_21_port_28_port_rcv_data     18648
D u64      LID_21_port_28_port_xmit_packets   242
D u64      LID_21_port_28_port_rcv_packets    259
D u64      LID_21_port_27_Remote_LID        14
D u64      LID_21_port_27_Remote_port       27
D u64      LID_21_port_27_port_xmit_data    39528
D u64      LID_21_port_27_port_rcv_data     24768
D u64      LID_21_port_27_port_xmit_packets   549
D u64      LID_21_port_27_port_rcv_packets    344
D u64      LID_21_port_26_Remote_LID        14
D u64      LID_21_port_26_Remote_port       26
D u64      LID_21_port_26_port_xmit_data    46955664
D u64      LID_21_port_26_port_rcv_data     27504
D u64      LID_21_port_26_port_xmit_packets   652162
D u64      LID_21_port_26_port_rcv_packets    382
D u64      LID_21_port_25_Remote_LID        14
D u64      LID_21_port_25_Remote_port       25
D u64      LID_21_port_25_port_xmit_data    2312115
D u64      LID_21_port_25_port_rcv_data     49247673
D u64      LID_21_port_25_port_xmit_packets   32156
D u64      LID_21_port_25_port_rcv_packets    683996
D u64      job_id                         533
```

SAMPLERS

- Using GraphViz and by selecting Remote LID and port, we can use the LID, Port->Remote LID, Remote Port connections to see fabric connections and identify any wiring issues.



OPENFABRICS ALLIANCE

INITIAL RESULTS

INITIAL RESULTS

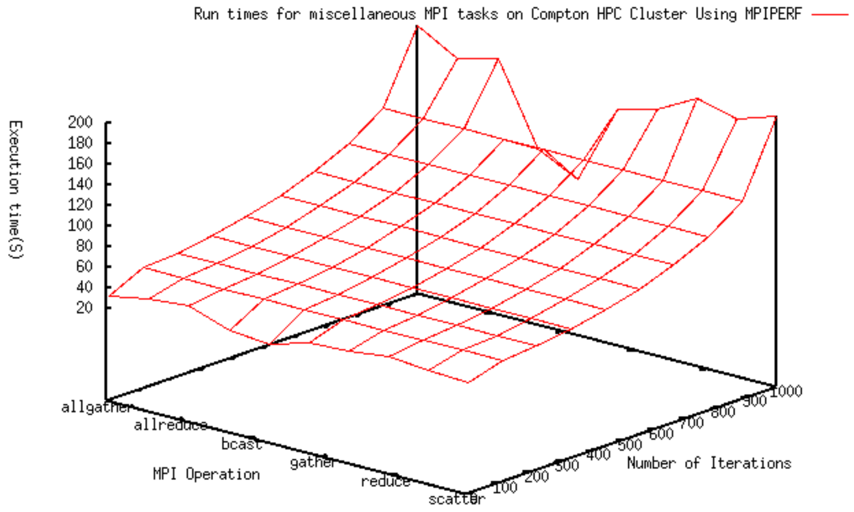
- **Initial Results**
  - An example of how long it takes to gather and process metrics from switch ports?
    - The sampler processes 0 metrics
    - The sampler processes 2 metrics
    - The sampler processes 4 metrics
    - The sampler processes a full set of available metrics
  - What is the load on the system with in-band metric gathering?
    - Tests run with simulated high MPI RDMA exchange traffic with in-band metric gathering.
  - Metrics retrieved by sampler.
    - Traffic metrics.
    - Congestion metrics

IBFABRIC TIME SPENT PROCESSING METRICS

- **MAD queries were made to switch ports**
- **Picking and directly connecting an HCA port to local switch ports makes a big difference in sampling performance.**
  - After retrieving metric information from the switch ports, 0 metric information was placed into an LDMS set for the Aggregator. Mean run-time for 100 metric gathers of the IBFabric Sampler was 1043 uS with a standard deviation of 175 uS. **The port was not connected directly to the sampled switch.**
  - After retrieving metric information from the switch ports, 2 metrics were placed into an LDMS set for the Aggregator. Mean run-time for 100 metric gathers of the IBFabric Sampler was 780 uS with a standard deviation of 165 uS. **The port was directly connected to the sampled switch**.
  - After retrieving metric information from the switch ports, 4 metrics were placed into an LDMS set for the Aggregator. Mean run-time for 100 metric gathers of the IBFabric Sampler was 702 uS with a standard deviation of 131 uS. **The port was directly connected to the sampled switch**.
  - After retrieving metric information from the switch ports, 24 metrics were placed into an LDMS set for the Aggregator. Mean run-time for 100 metric gathers of the IBFabric Sampler was 743 uS with a standard deviation of 191 uS. **The port was directly connected to the sampled switch**.

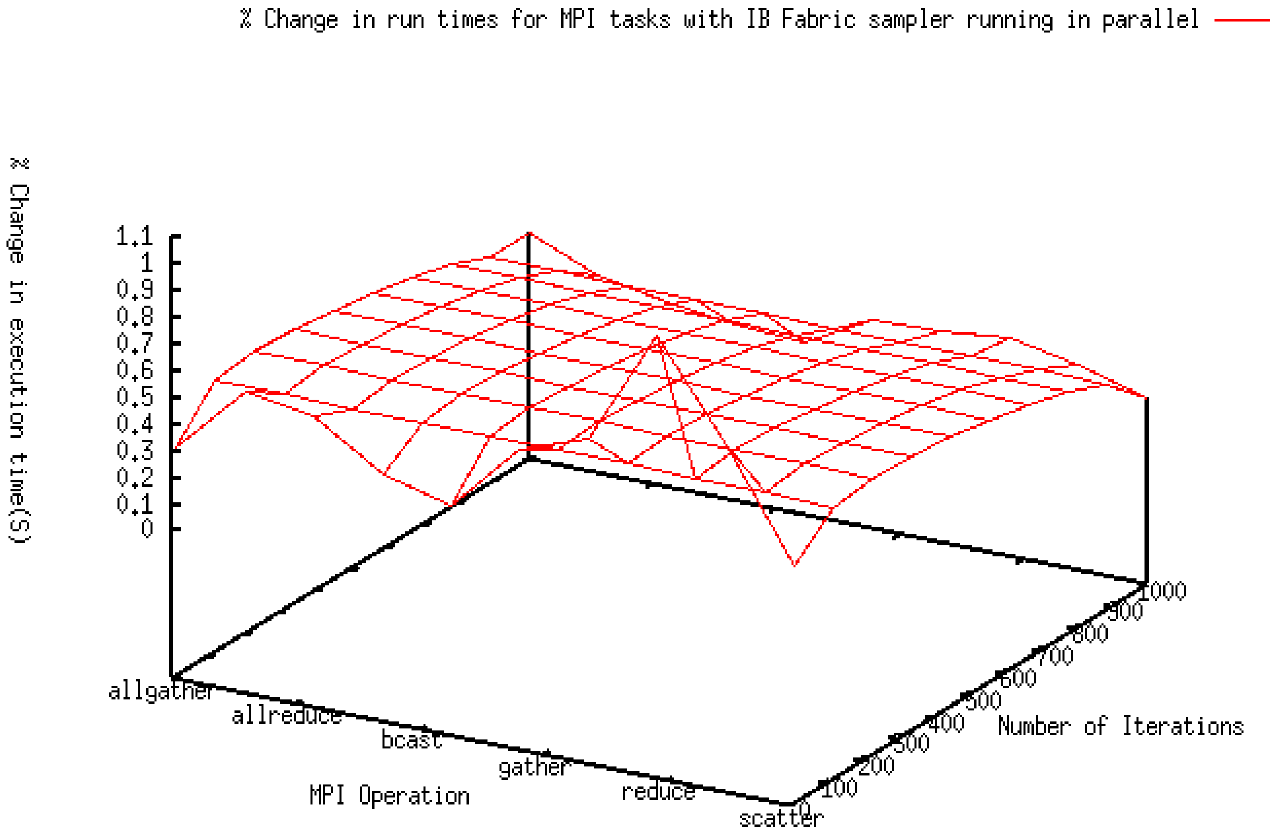
IN-BAND METRIC GATHERING OVER VL15

- **When MAD queries are made, metrics from the switch ports are higher priority than data exchanges. MAD requests are done on VL15**
- **How did that affect computations that required high levels of RDMA data exchanges?**
  - We simulated computations that performed generated high levels of MPI traffic to see the results



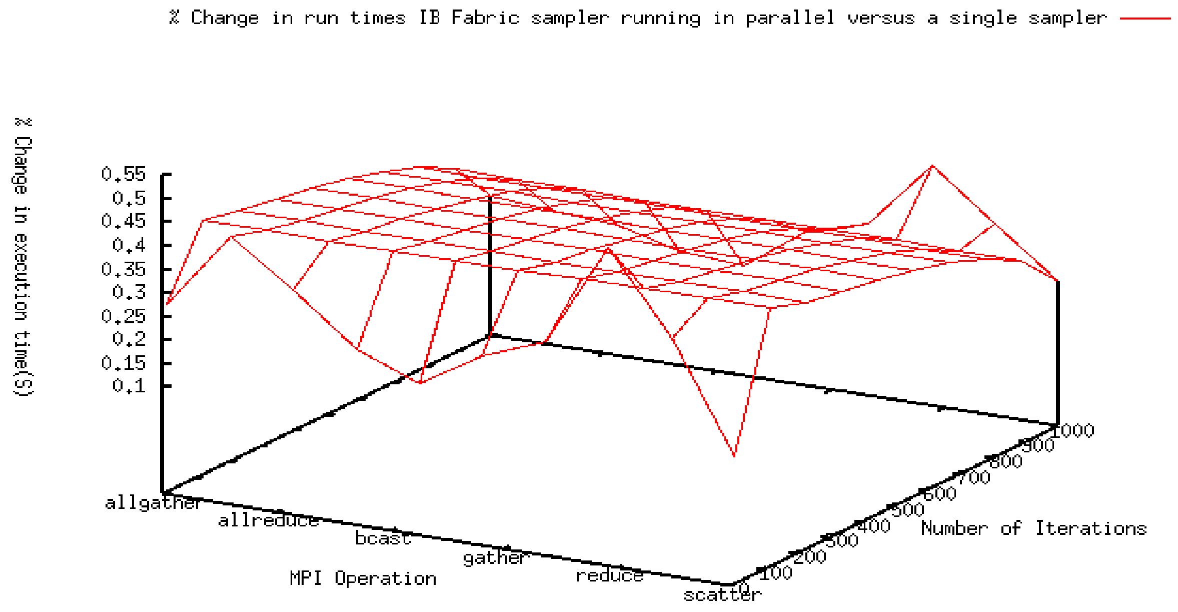
MPIperf Test Runs without IBFabric Sampler Running.

IN-BAND METRIC GATHERING OVER VL15



Percentage Change in MPIperf Run Times with
IBFabric Sampler Running in Parallel on Each Compute Node.

IN-BAND METRIC GATHERING OVER VL15



Percentage Change in MPIperf Run Times with IBFabric Sampler Running in Parallel on Each Compute Node versus IBFabric Sampler Running on a Single Compute Node and Reading the Entire Fabric.

UPGRADES AND FUTURE WORK

UPGRADES AND FUTURE WORK

- **Currently, if a sampler goes down, the metric gathering for a group of switch ports is lost**
  - We might like a way to notify the other samplers to be notified that a sampler is down and to begin sampling switch ports inn the absence of a lost sampler.
  - Alternatively, we can collect data from each switch port using N sampling hosts to provide n-way redundancy and filter after to eliminate duplicates.
- **Reduction in the number of switch samples by a factor of 2**
  - With more advanced logic, we will be able to reduce the number of samples by 2 due to the fact that we will only sample one end of a connection instead of both ends.
  - Lower overhead, but less robust to failure.
- **Create a fabric sampler to work with OmniPath.**

CONCLUSION

CONCLUSION

- **New measurement tool developed.**
- **Demonstrated low overhead on test cluster hardware.**
- **Reveals directly the previously inaccessible data switch behavior.**

OPENFABRICS ALLIANCE

13th ANNUAL WORKSHOP 2017

THANK YOU

Michael Aguilar

Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2013-6521C