14th ANNUAL WORKSHOP 2018

COMPREHENSIVE, SYNCHRONOUS, HIGH FREQUENCY MEASUREMENT OF INFINIBAND NETWORKS IN PRODUCTION HPC SYSTEMS

Benjamin Allan, Michael Aguilar, and Serge Polevitzky

[April, 2018]

U.S. DEPARTMENT OF ENERGY | NNSA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
SAND No. 201X-XXXX.

Sandia National Laboratories

SAND Number: SAND2018-1834 C

SAND2018-3506C

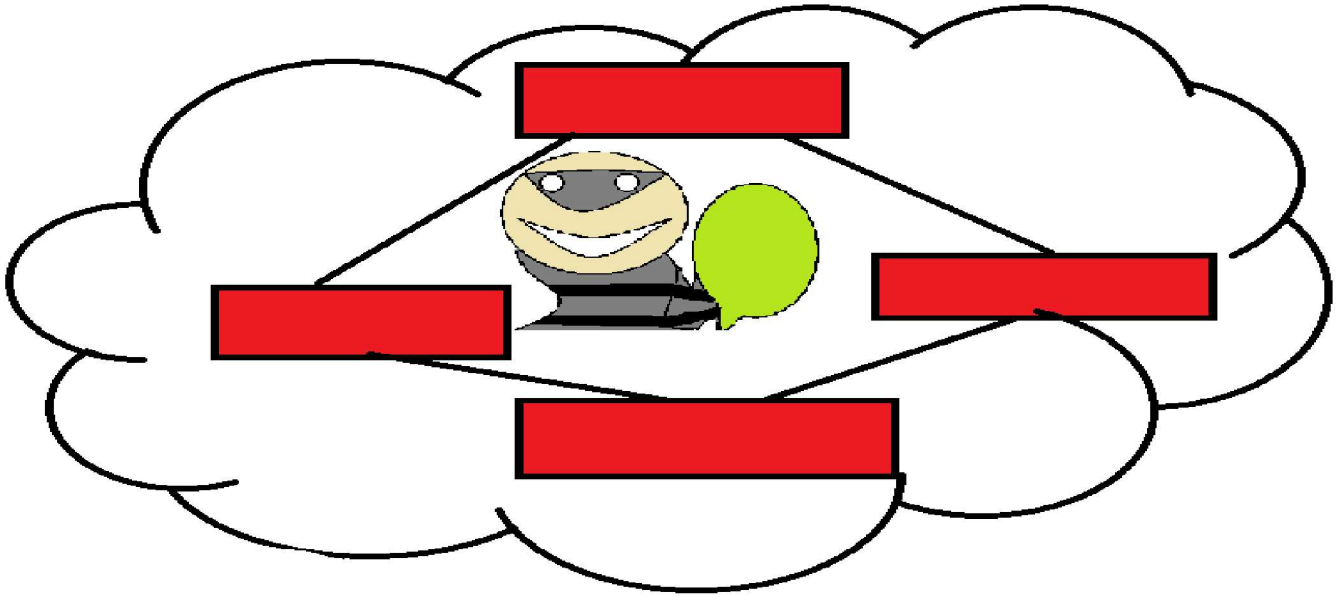
OUTLINE

- **Why Synchronous Performance Data Gathering?**
- **Challenge of Pulling Performance Data from Big HPC Fabrics**
  - How Previous Experience Shaped Our Approach
- **Experiments**
- **Results**
- **Conclusions**
- **Questions?**

WHY DO WE CARE ABOUT STRICT SAMPLE INTERVALS?

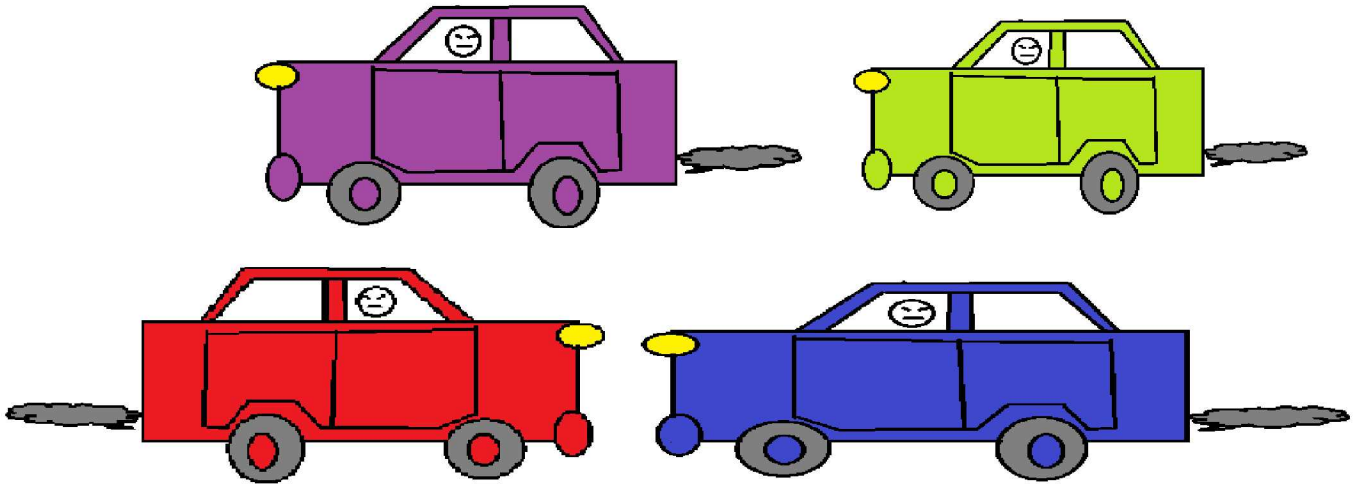
STRICT SAMPLING INTERVALS

- **We are looking for network related issues that slow our computational performance.**



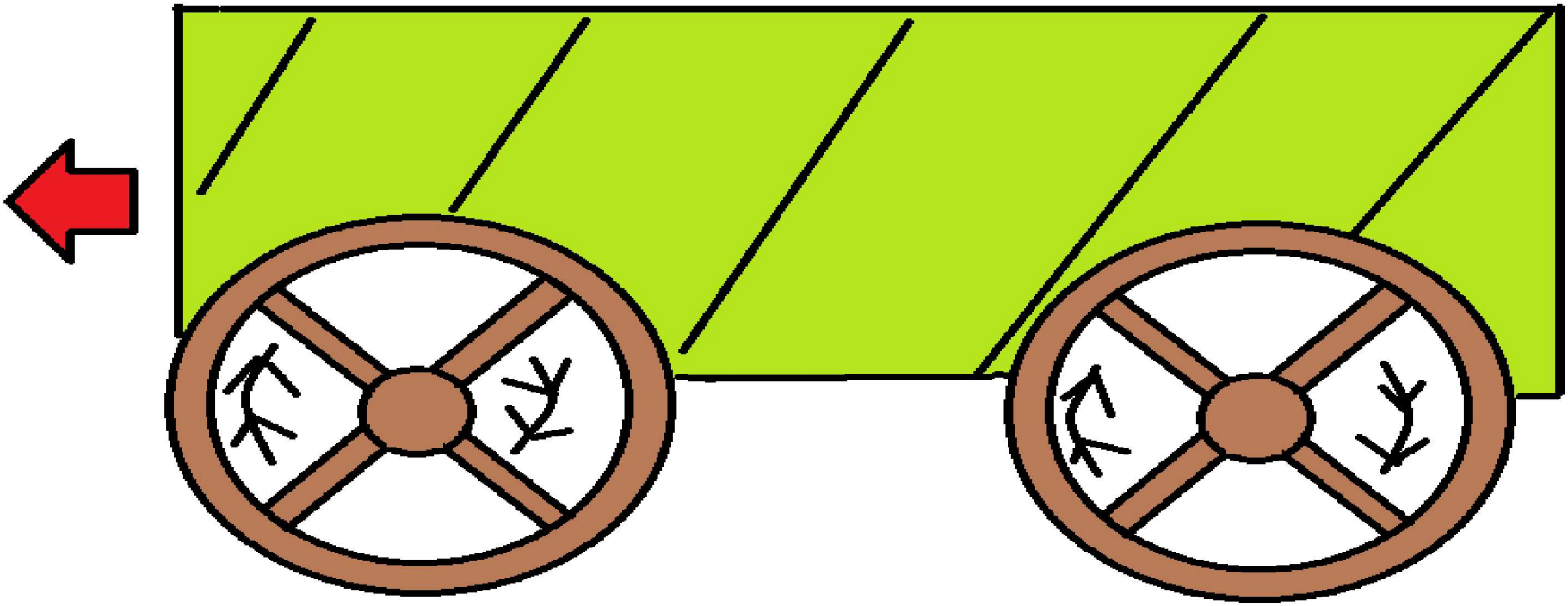
STRICT SAMPLING INTERVALS

- Easier to correlate performance data with running jobs.
- Shorter sampling intervals allow us to more easily see dynamic changes in network traffic.



STRICT SAMPLING INTERVALS

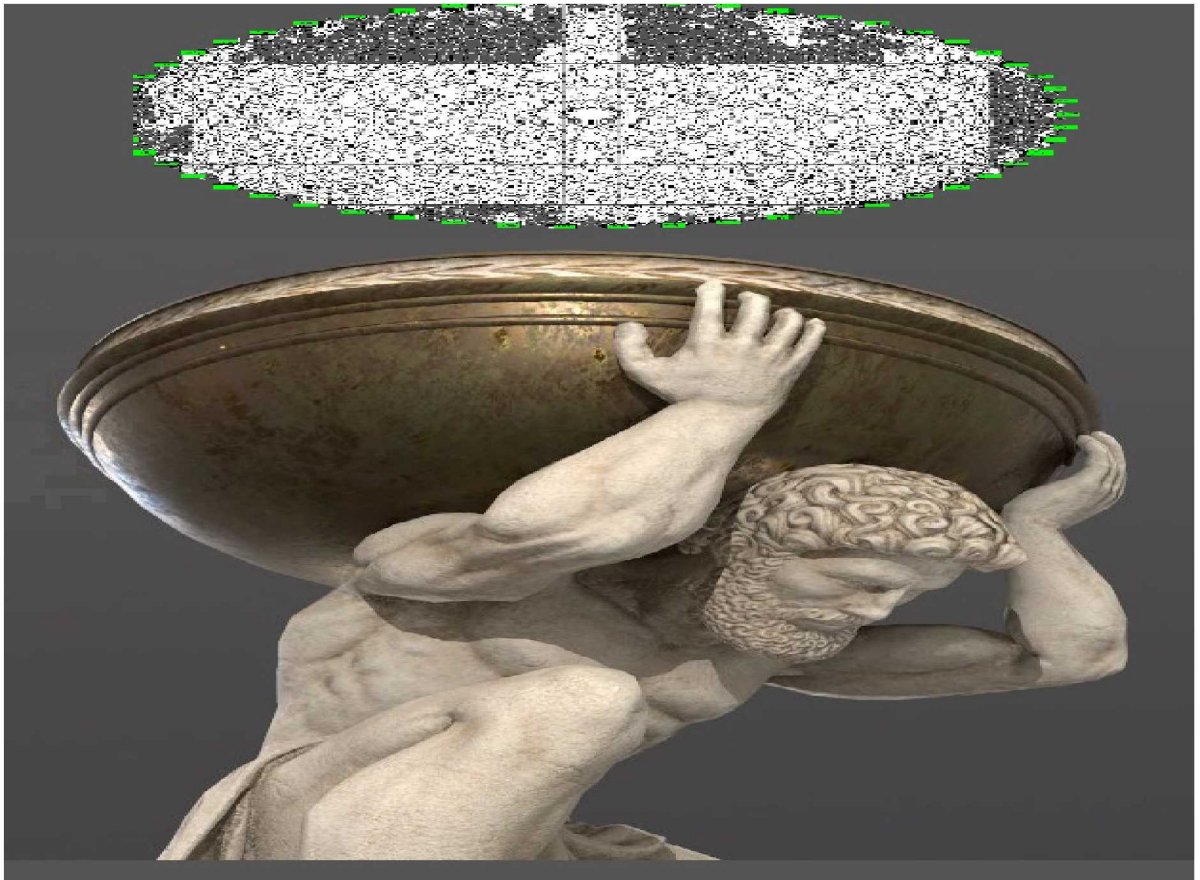
- The more infrequently we gather performance statistics, the more we smooth away information. Peaks get hidden.



CHALLENGE OF PULLING PERFORMANCE DATA FROM BIG HPC FABRICS

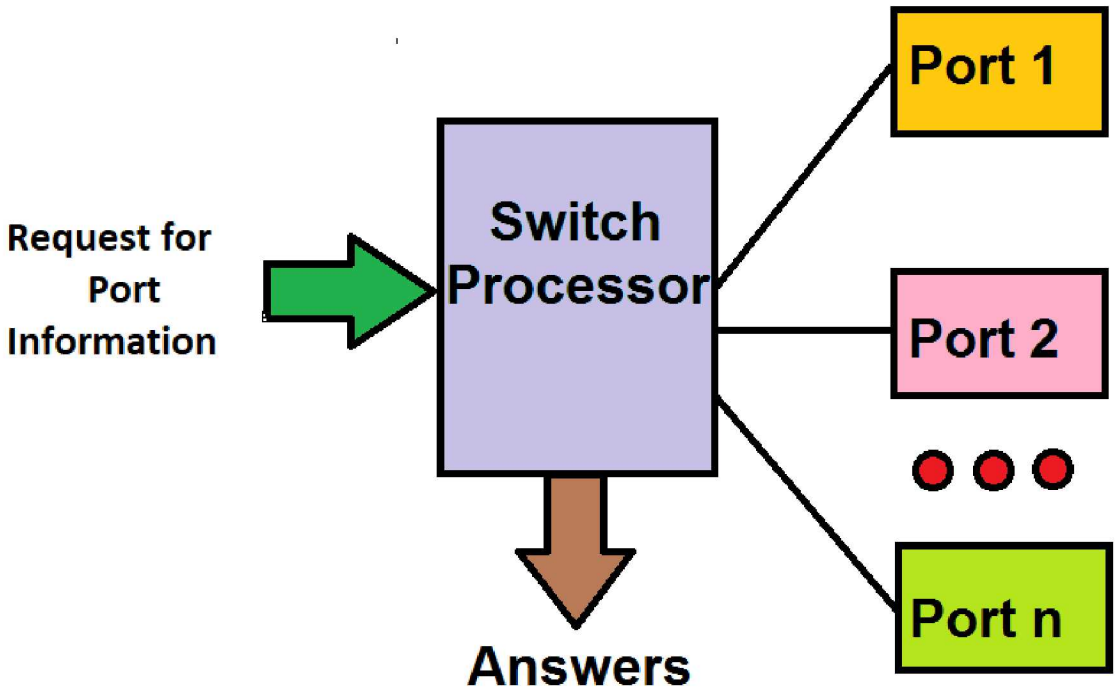
CHALLENGE OF PULLING DATA FROM BIG HPC FABRICS

- We want to create the least interference to network traffic for running applications.
- We want the minimal retrieval time for our queries.

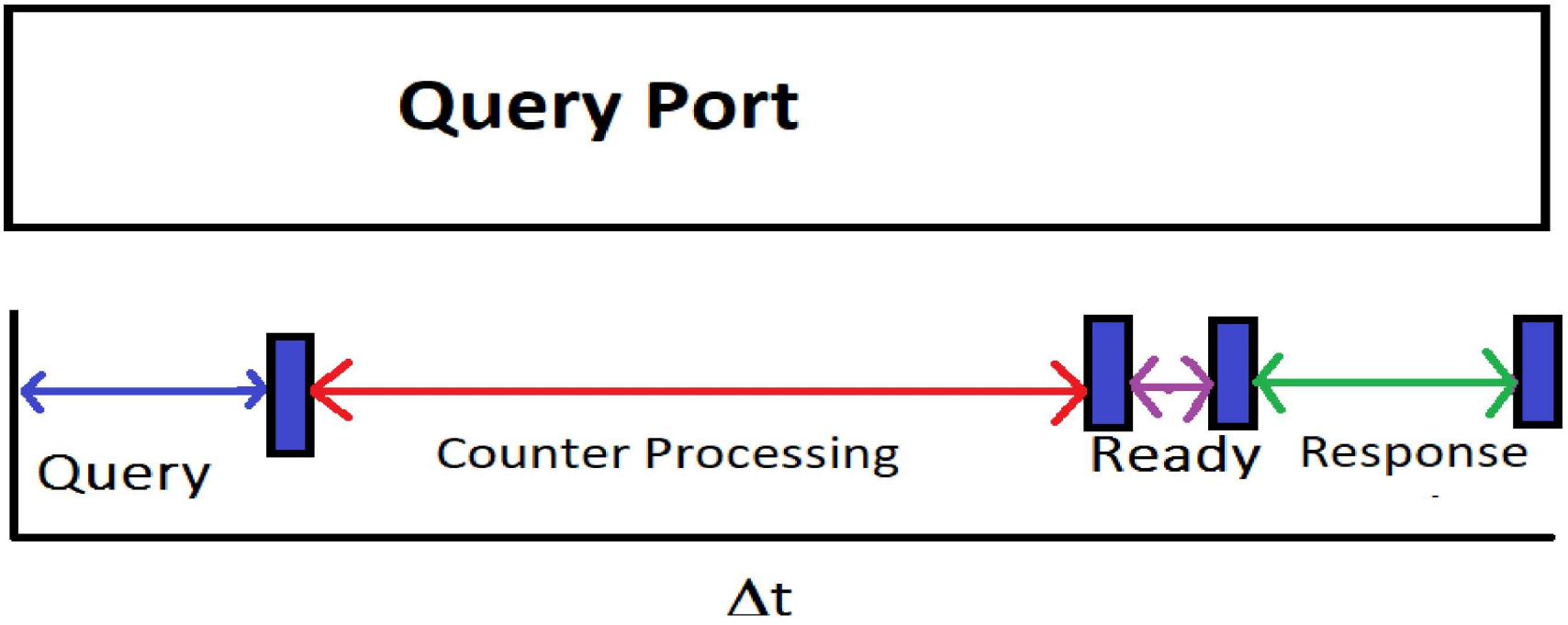


CHALLENGE OF PULLING PERFORMANCE DATA FROM BIG HPC FABRICS

- **Requests are made to the switch and then the switches retrieve performance metric data.**

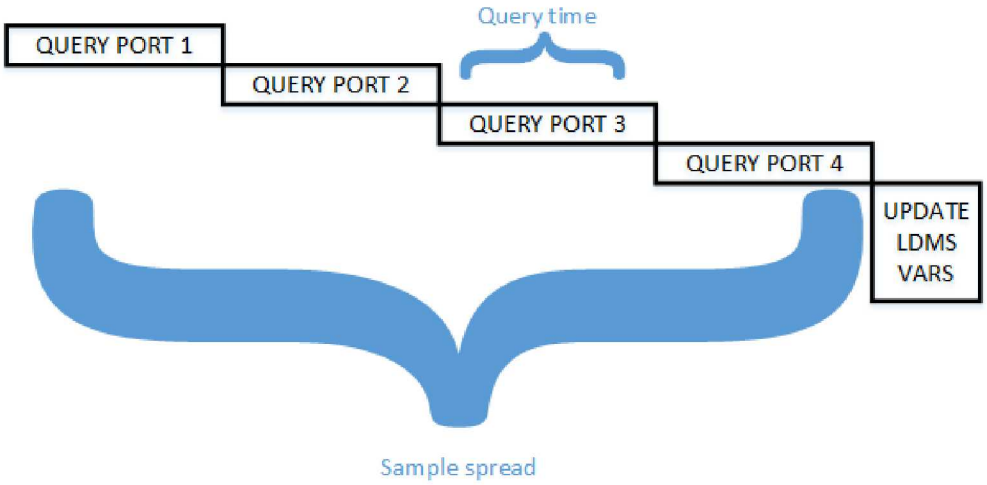


CHALLENGE OF PULLING PERFORMANCE DATA FROM BIG HPC FABRICS



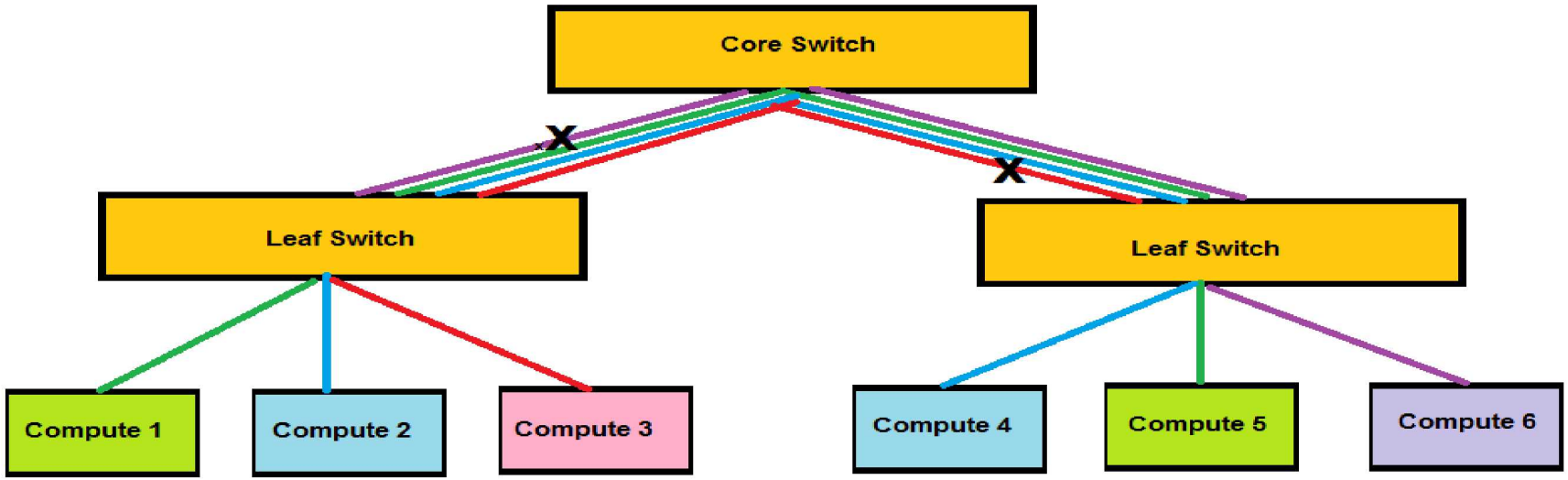
CHALLENGE OF PULLING DATA FROM BIG HPC FABRICS

- **For a sampler on a single node serially querying a list of IB ports, we define spread as the time interval between the start of the first port query made and the end of the last port query made.**



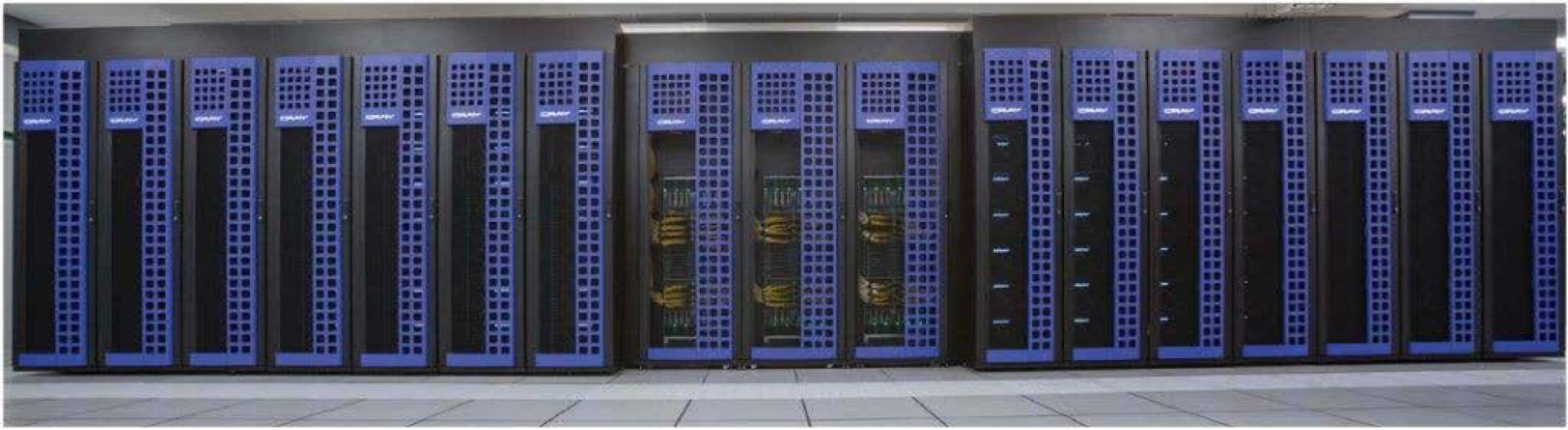
CHALLENGE OF PULLING DATA FROM BIG HPC FABRICS

- Previous experiments showed us that gathering switch port information from closely connected samplers is better than having queries traverse the fabric to a far switch.
- Previous experiments showed us that up to 1 Hz sampling rates do not negatively affect application traffic.



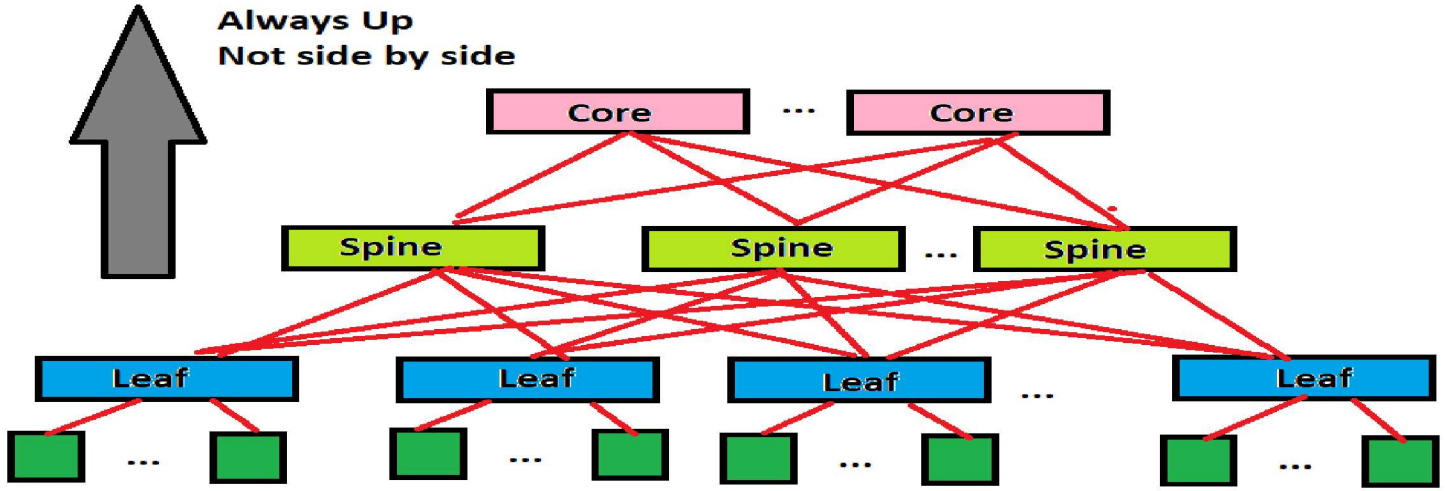
EXPERIMENTS

- **Skybridge----Top 500 rank of 381, 1848 Compute Nodes (6/2017)**



EXPERIMENTS

- A very large 3-tier Fat Tree
- Sampling in our demonstration is performed using Skybridge Administrative Nodes.
- 268 Switches, 9648 Switch Ports



RESULTS

RESULTS

- **Individual port query time statistics for 10 Samplers on Skybridge:**

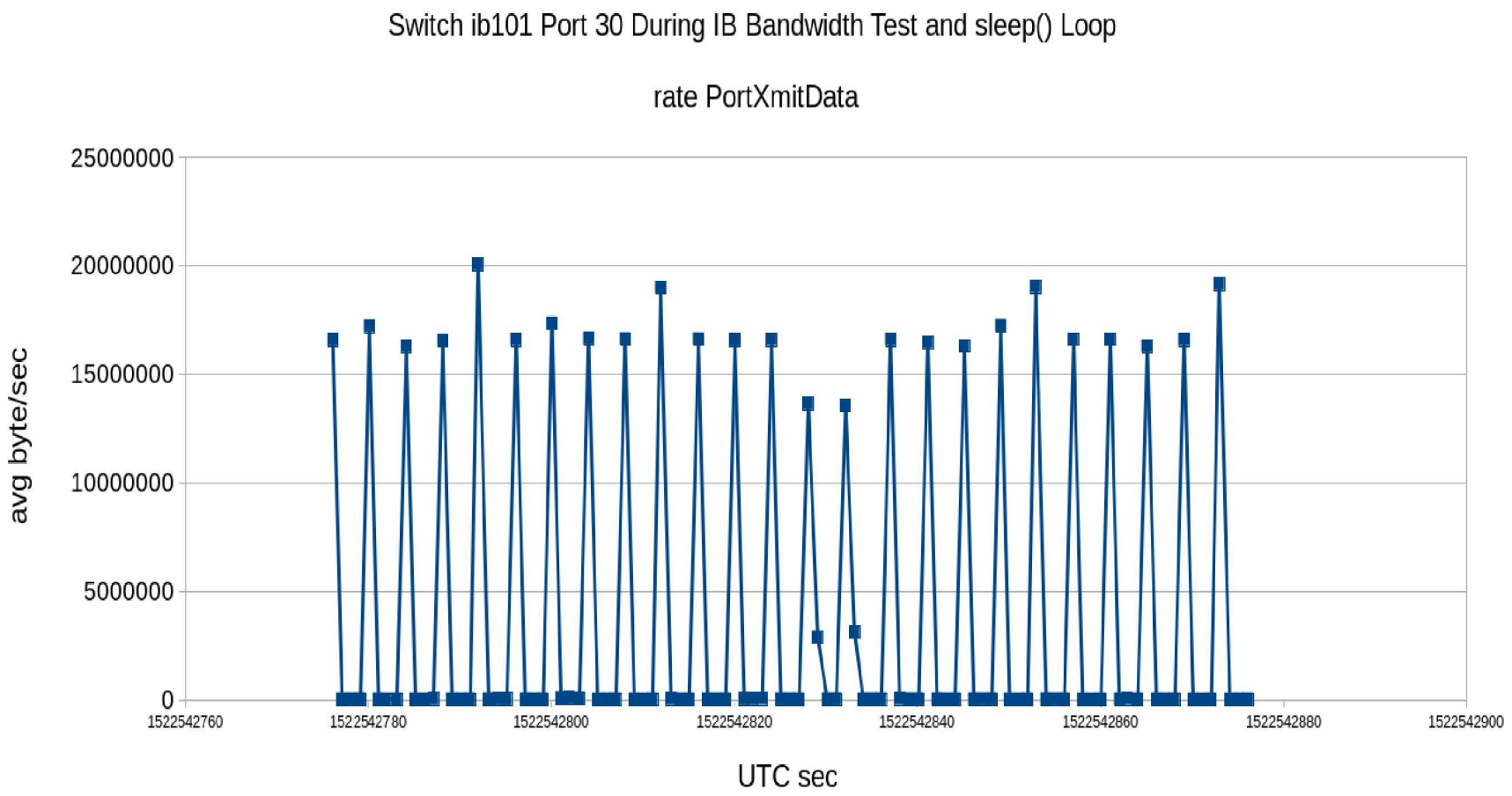
  - Retrieval time for each switch port
    - avg 0.00014
    - min 0.000048
    - max 0.013 **<== 75 Hz maximum practical sampling frequency.**

  - Time for a sampler to collect its share of ports:
    - min 0.105
    - max 0.224
    - avg 0.149

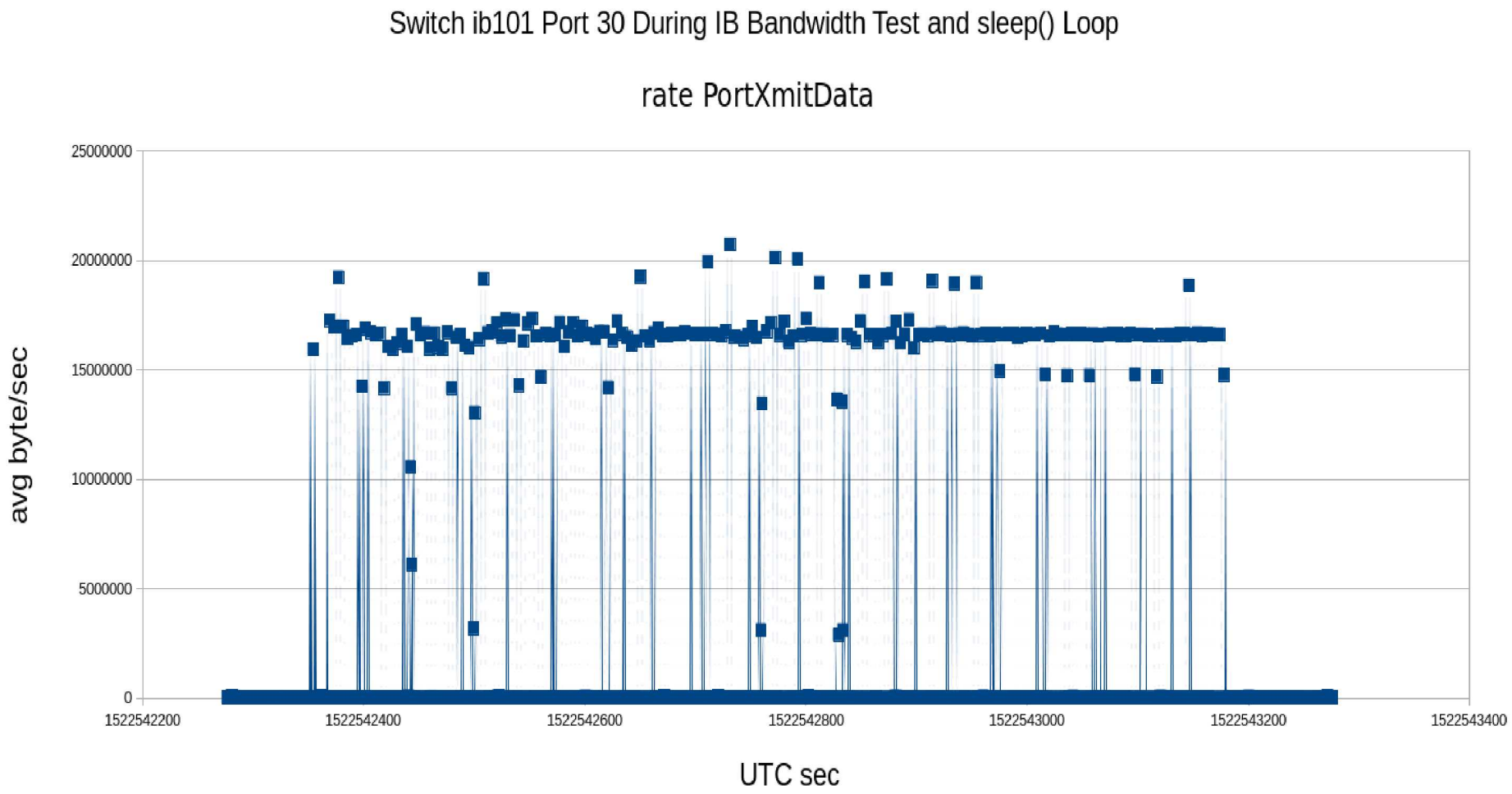
RESULTS

- **Port query time statistics for 1 Sampler on Skybridge**
  - Retrieval time for each switch port
    - avg 0.00065 s
    - min 0.000074 s
    - max 0.0038 s
  - Time for a single sampler to collect all ports on all switches:
    - min 6.05 s
    - Max 6.42 s
    - avg 6.17 s
- **No IB errors were detected during the tests (we checked).**

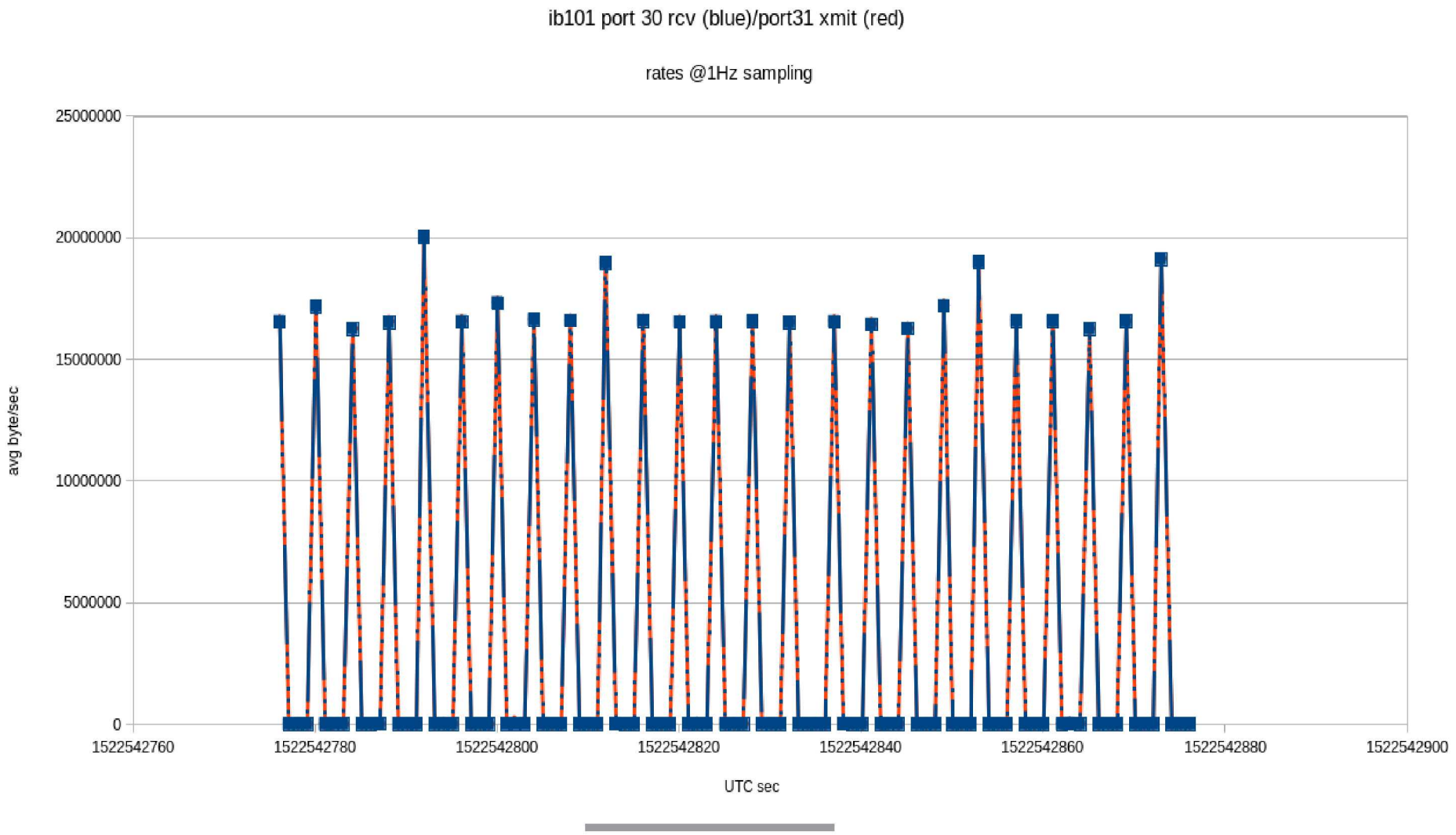
RESULTS (3 MINUTES)



RESULTS (30 MINUTES)



RESULTS (2 MINUTES)



CONCLUSIONS AND ACKNOWLEDGMENTS

CONCLUSIONS

- **We can scan the entire fabric on a large system in fixed time intervals.**
  - A Single IBFabric Sampler running on a Single Skybridge Admin node, sampling all of the Skybridge InfiniBand switches can be done every 20s
  - When we have samplers running concurrently on 10 Skybridge Admin nodes, sampling can be done at 1Hz.
  - If we sample one switch per compute node, sampling can be done at 10 Hz.
- **We were able to see network performance data for our test traffic.**
- **We can sample a full suite of performance and error metrics from the switches without inducing errors.**
- **We saw VL15 drops on the Slurm and OpenSM node on ~1 minute periodic basis. *Which service that is running in the background is causing this?***

ACKNOWLEGEMENTS

Steve Monk, Mark Schmitz, Joe Mervini

QUESTIONS?

14th ANNUAL WORKSHOP 2018

THANK YOU

Benjamin Allan, Michael Aguilar, Serge Polevitzky

Sandia National Laboratories

SAND Number: SAND2018-1834 C