OPENFABRICS ALLIANCE

14th ANNUAL WORKSHOP 2018

LDMS AND INFINIBAND @ SANDIA

Serge Polevitzky, SAIC HPC Support at SNL
Mike Aguilar, Ben Allan, Char Arias, Jay Livesay, Justin Wood, and Many Others

[April, 2018]

Sandia National Laboratories

SAND Number: SANDXXX-YYY

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.

INTRODUCTION

- **What You Can Expect Today**
  - **Status of LDMS IB Fabric Investigation**
  - **Peeks Behind the LDMS Curtain**
  - **Background on How the LDMS Tools Used**
  - **Stumbles, Fumbles, & Recoveries**

GETTING, VERIFYING, UNDERSTANDING "THE DATA"

ASSUMPTIONS & SOME GOALS

- **1) We Have A Hard Enough Time Just Keeping Clusters Up**
- **2) We Haven't Had ~~Much~~ *Any* Time to Look Deep into Our Fabrics, Let Alone Our IB Tools**
- **3) Goal Today Is 1st to Validate What We *Think* We Know, 2nd to Look at Some LDMS Fabric Data, and 3rd to Look at Things We Can Infer from the Data**
- **4) Finally, Draw from 3) Above & Provide Insight for Ourselves (Technical Folk) and for Management**

GROUND RULES

- **What Follows Is Not …**
- **… *Not an LDMS v. other tools* Rant (far from it)**
- **"other tools" & LDMS Benefit from *Each Other***
  - ✓ **"One Hand Washes the Other"**
  - ✓ **One Tool Validates the Other,**
  - ✓ **LDMS Provides 2nd Opinion for "other tools"**
  - ✓ **"other tools" Provide 2nd Opinion for LDMS**

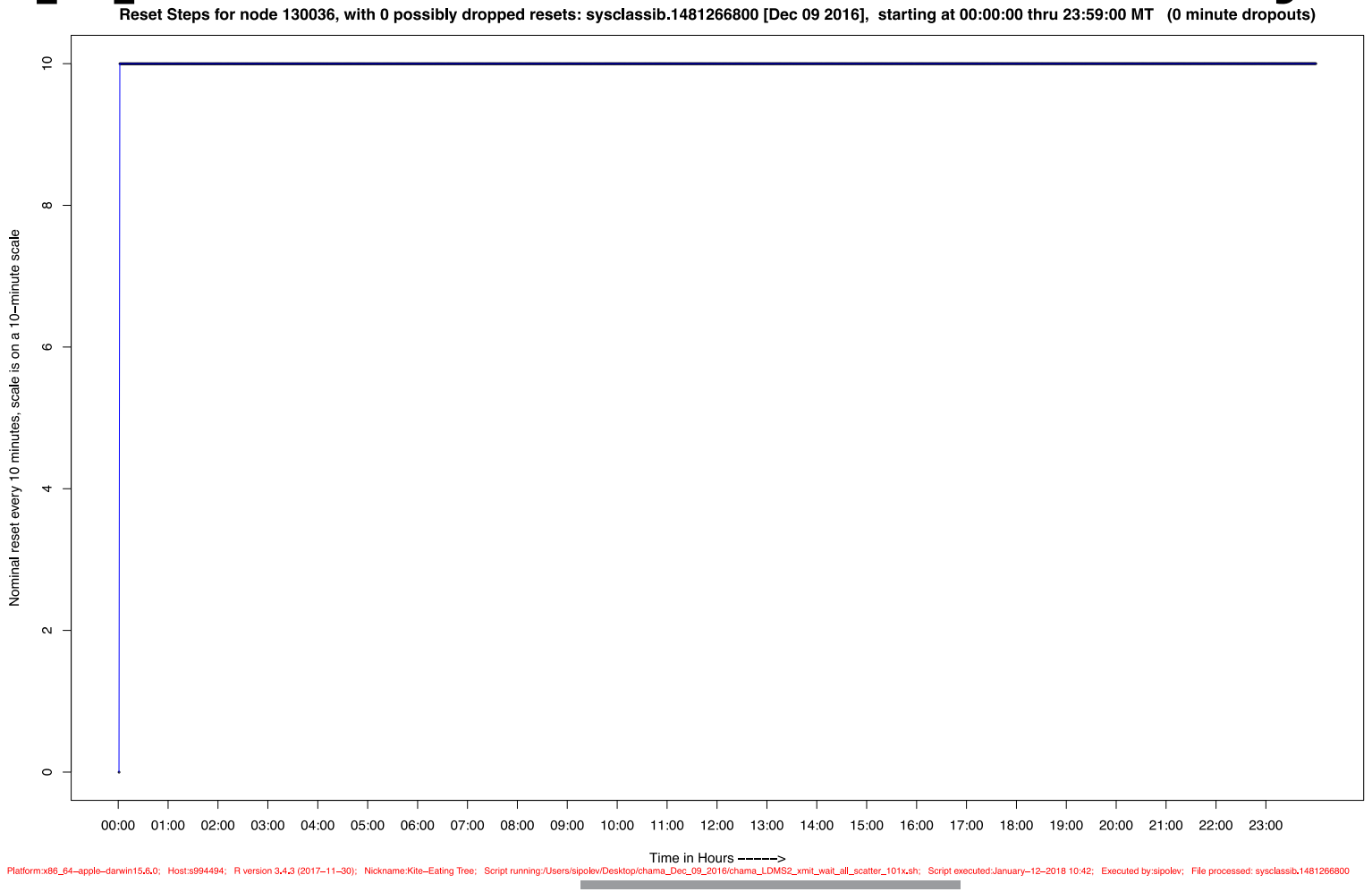
TRYING TO VALIDATE WHAT WE ALREADY "KNOW")

- **At Sandia: *IBMON* Broadcast Resets to HCAs Every 10 minutes***
- **We Should See These Resets in the LDMS IB HCA Data**
- **Resets Only Visible If Non-Zero Values Before a Reset****
- **Let's Look at Some Data & Does It Verify What We Expect ?**

*** via calls to ib_counters.sh from ibmon.pl**

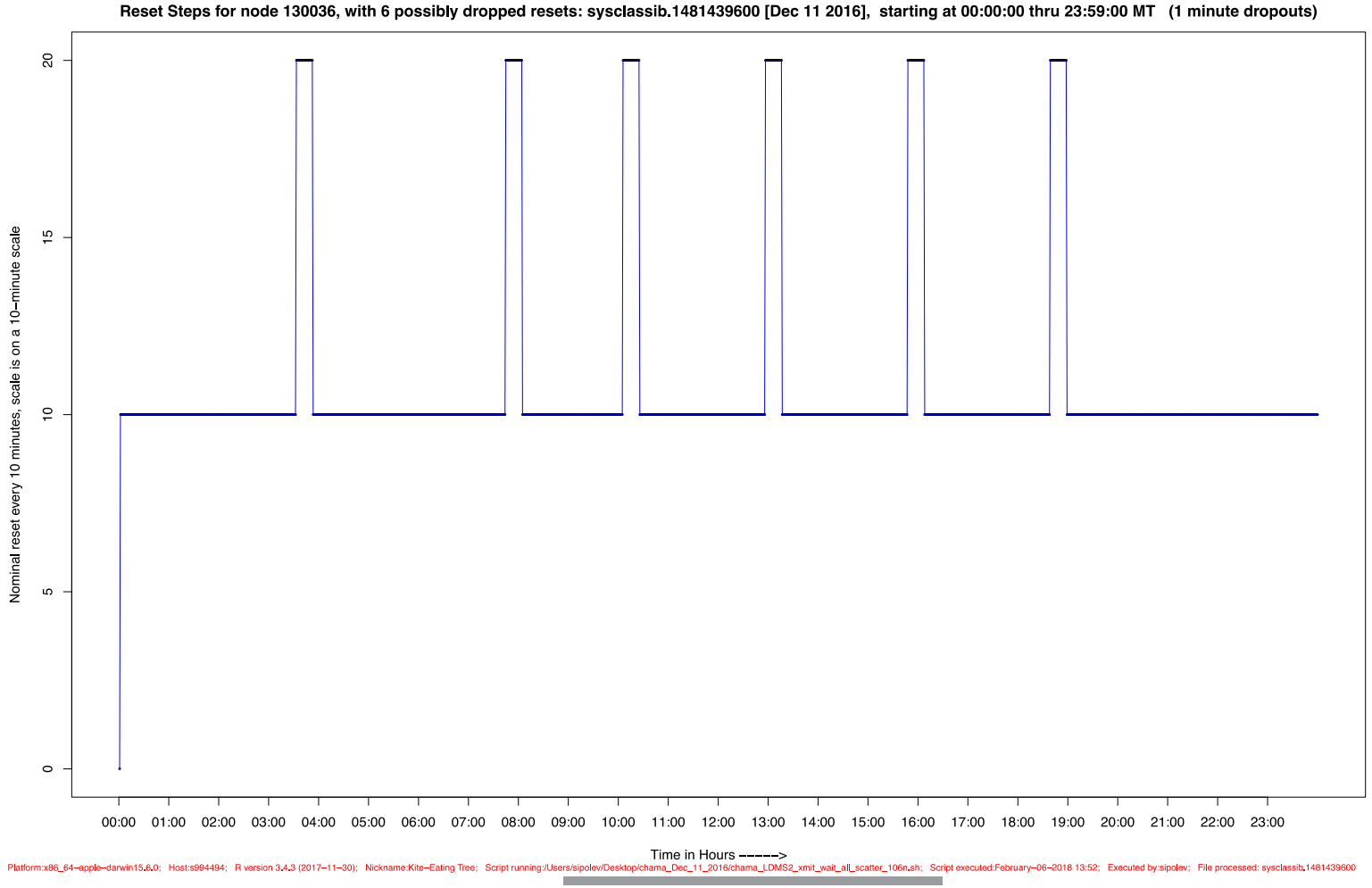
RESETS !

- **Exhibit [A]: JOY ! PTW Counts Reset Every 10-Minutes !**



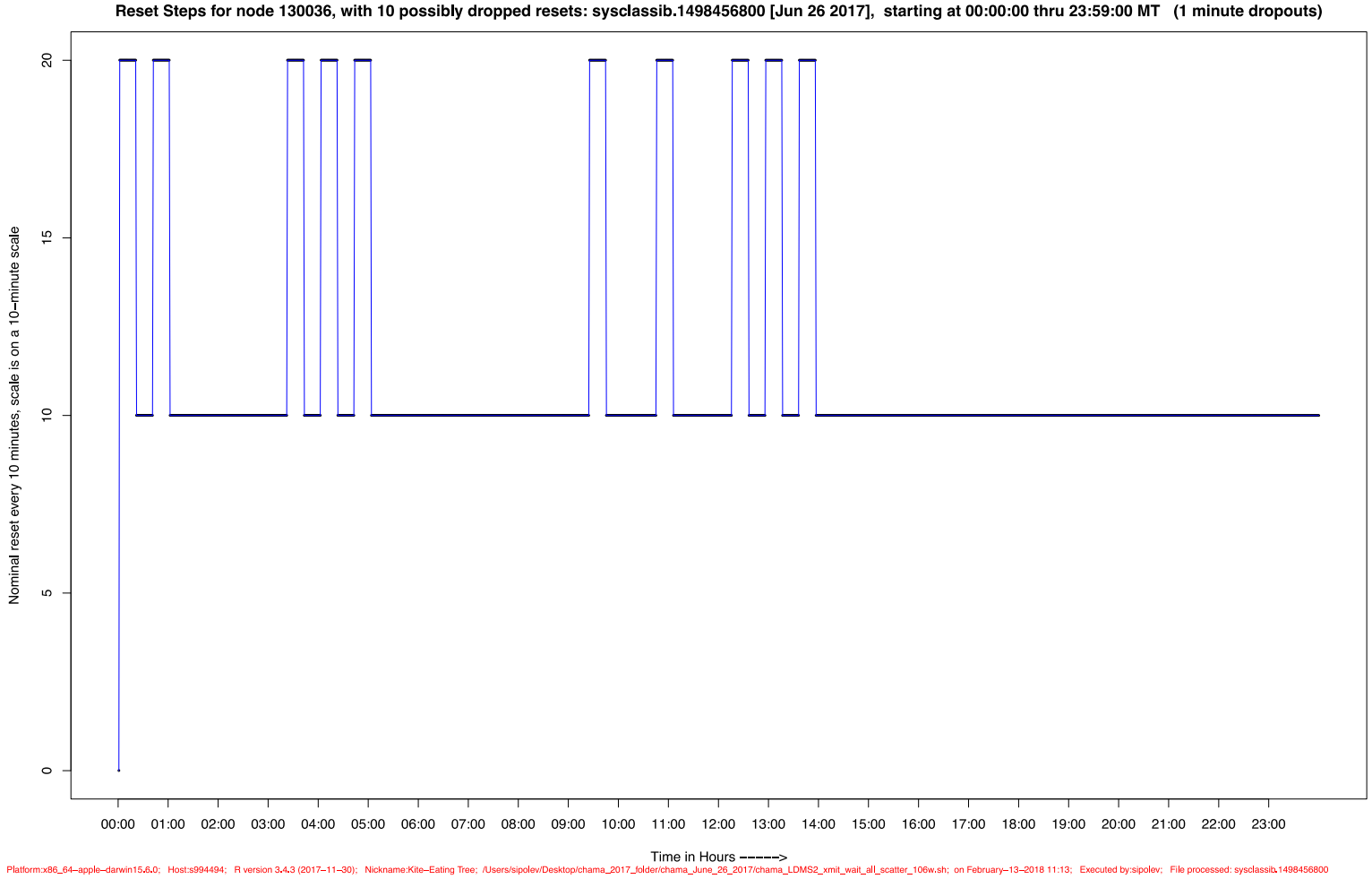
RESETS !!

- **Exhibit [B]: Oops, PTW Counts Same Node — *Not* Every 10 Minutes**

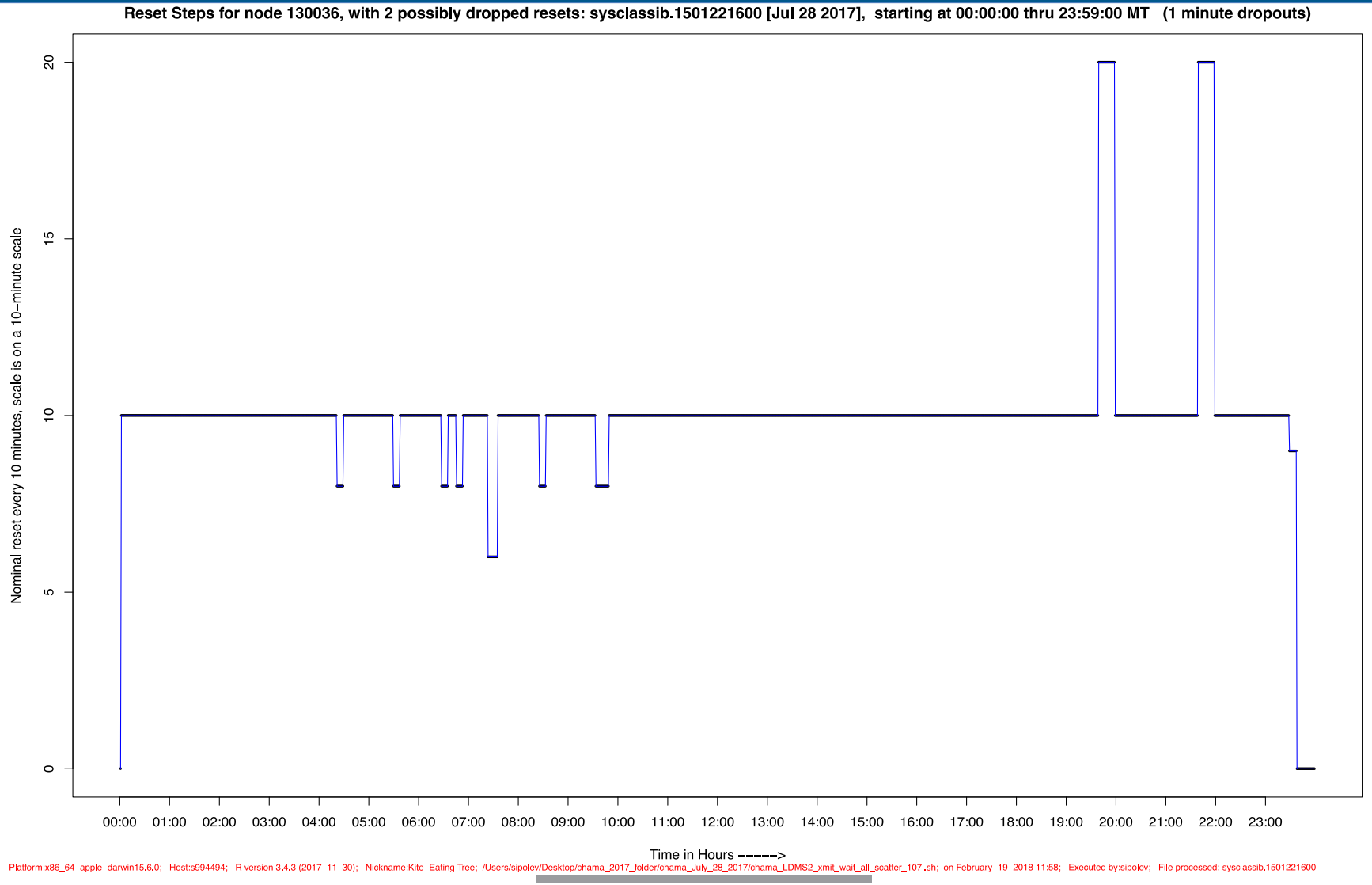


RESETS MORE SPORADIC

- **Exhibit [C]: Same Node “Resets Not Every 10 Minutes”**



RESET CYCLES MAYBE SOMETIMES “CHAOTIC”



RESETS DROPPED NOT JUST FROM PTW COUNTS

Exhibit [D]:

Gateway 48	Unicast Packet Xmit	Delta		receive_data	Delta	xmit_data	Delta
June 26, 2017							
130048	3.35E+16			8.16628E+08		2.00668E+09	
130048	2.7785E+14	-3.32221E+16		2.20881E+07	-7.94539E+08	4.63833E+07	-1.96029E+09
130048	9.81838E+14	7.03988E+14		1.42974E+08	1.20886E+08	2.62807E+08	2.16424E+08
130048	1.39328E+15	4.11445E+14		2.10161E+08	6.71866E+07	3.93935E+08	1.31128E+08
130048	1.61074E+15	2.17454E+14		2.62346E+08	5.21851E+07	4.55734E+08	6.17988E+07
130048	1.77679E+15	1.66056E+14		3.73782E+08	1.11436E+08	5.10090E+08	5.43562E+07
130048	1.97122E+15	1.94425E+14		3.84075E+08	1.02934E+07	5.72428E+08	6.23386E+07
130048	2.45962E+15	4.88407E+14		3.93668E+08	9.59300E+06	6.71085E+08	9.86565E+07
130048	2.73741E+15	2.7779E+14		4.04760E+08	1.10913E+07	9.03292E+08	2.32207E+08
130048	2.76969E+15	3.22767E+13		4.47098E+08	4.23388E+07	9.59613E+08	5.63212E+07
130048	2.78451E+15	1.48219E+13		5.88688E+08	1.41590E+08	1.00448E+09	4.48671E+07
130048	2.65488E+16	2.37643E+16		1.84430E+08	-4.04258E+08	5.56857E+08	-4.47623E+08
130048	2.67988E+16	2.49941E+14		5.66980E+08	3.82550E+08	6.66753E+08	1.09896E+08
130048	2.69692E+16	1.70493E+14		7.56698E+08	1.89718E+08	7.82202E+08	1.15449E+08
130048	2.69973E+16	2.80504E+13		8.05137E+08	4.84383E+07	8.90086E+08	1.07884E+08
130048	2.71167E+16	1.1937E+14		8.98208E+08	9.30708E+07	1.07180E+09	1.81718E+08
130048	2.72235E+16	1.06786E+14		9.94729E+08	9.65211E+07	1.18068E+09	1.08881E+08
130048	2.78388E+16	6.15396E+14		1.01539E+09	2.06591E+07	1.29172E+09	1.11035E+08
130048	3.41591E+16	6.3202E+15		1.06713E+09	5.17419E+07	1.49450E+09	2.02785E+08
130048	3.4882E+16	7.22899E+14		1.35174E+09	2.84613E+08	1.69106E+09	1.96551E+08
130048	3.95811E+16	4.69917E+15		1.56717E+09	2.15426E+08	2.13116E+09	4.40108E+08
130048	2.9579E+14	-3.92853E+16		1.31546E+08	-1.43562E+09	1.52552E+08	-1.97861E+09
130048	4.52707E+14	1.56917E+14		2.44884E+08	1.13338E+08	2.66495E+08	1.13943E+08

HOW ARE RESETS DELIVERED ? BY VL15s ….

- **Recall That VL15s (Unlike VL0 – VL14) Are “Fire & Forget”**
- **User (VL0 – VL14) Requires “Credit Info” to Be Exchanged**
- **Again, VL15 Transactions Do Not …**
- **Other Management Datagrams (MADs) Also Go via VL15s:**
  - ***send -- \"perfmgr dump_counters\r\"***
  - ***expect \"*OpenSM*\”***
  - ***send -- \"perfmgr clear_counters\r\”***
- **Reads & Clears and More CRITICAL VL15s Can Be Dropped**
- **And, YES, Dropped VL15s Have Been of Concern ... Forever**

MUSING ON VL15s BEING DROPPED ….

- **Perhaps "Interesting" that Reset VL15s Are Being Dropped**
- **Maybe Dropped Reset VL15s "Not Important" (Unless Sums)**
- **Assume that the Sender Knows about Silent Reset VL15 Drops**
- **However, What About Other VL15s, VL15s that Are *Critical***

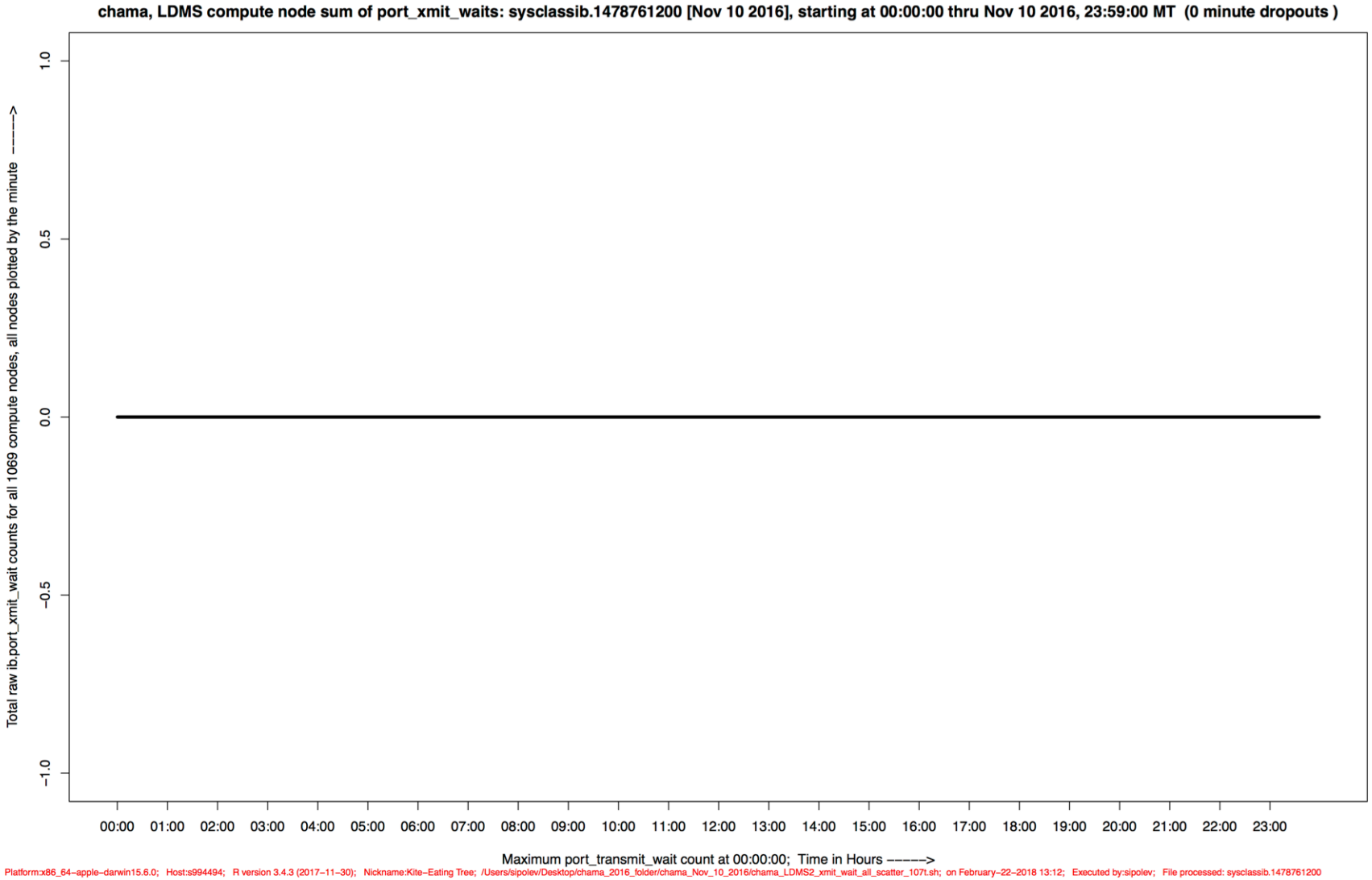
- **Segue: Consider the Law of Large Numbers**
- **Even With a Very Low "Drop Rate" As the # of Nodes Goes Up**
  - **Probability of *Critical* VL15s Getting Dropped Goes toward Certainty**
  - **Probability of a Successful Mitigation for Dropped VL15s Lowers**
- **Food for Thought for Very Large Clusters & For Exascale …**
- **End Segue**

GOING BACK TO PORT TRANSMIT WAIT COUNTS

- **Why ? PTW == Our Canary in Congestion Coal Mine**
  - **And Almost No Tool Looks at PTW Counts**
  - **Check Defaults (Probably PTW Captures Are OFF)**
  - **Remember: PTW Counts "Rail" [32-bit counters]**
- **PTWs Indicate Subtle Problems {Unlike Symbol Errors, Link Downs, Port Receive Errors, Which Are 'Obvious'}**

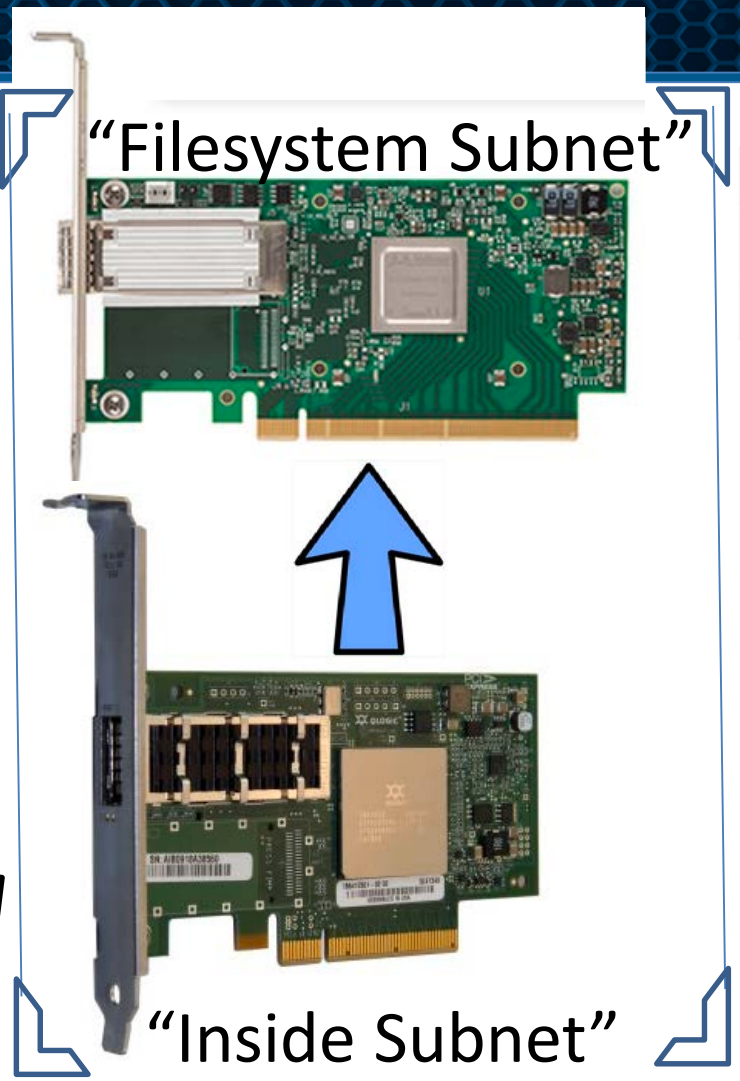
WHERE DO WE FIND PTW COUNTS ?

Exhibit F



PORT TRANSMIT WAIT COUNTS ONLY ON GATEWAYS

Exhibit G



Gateway Node

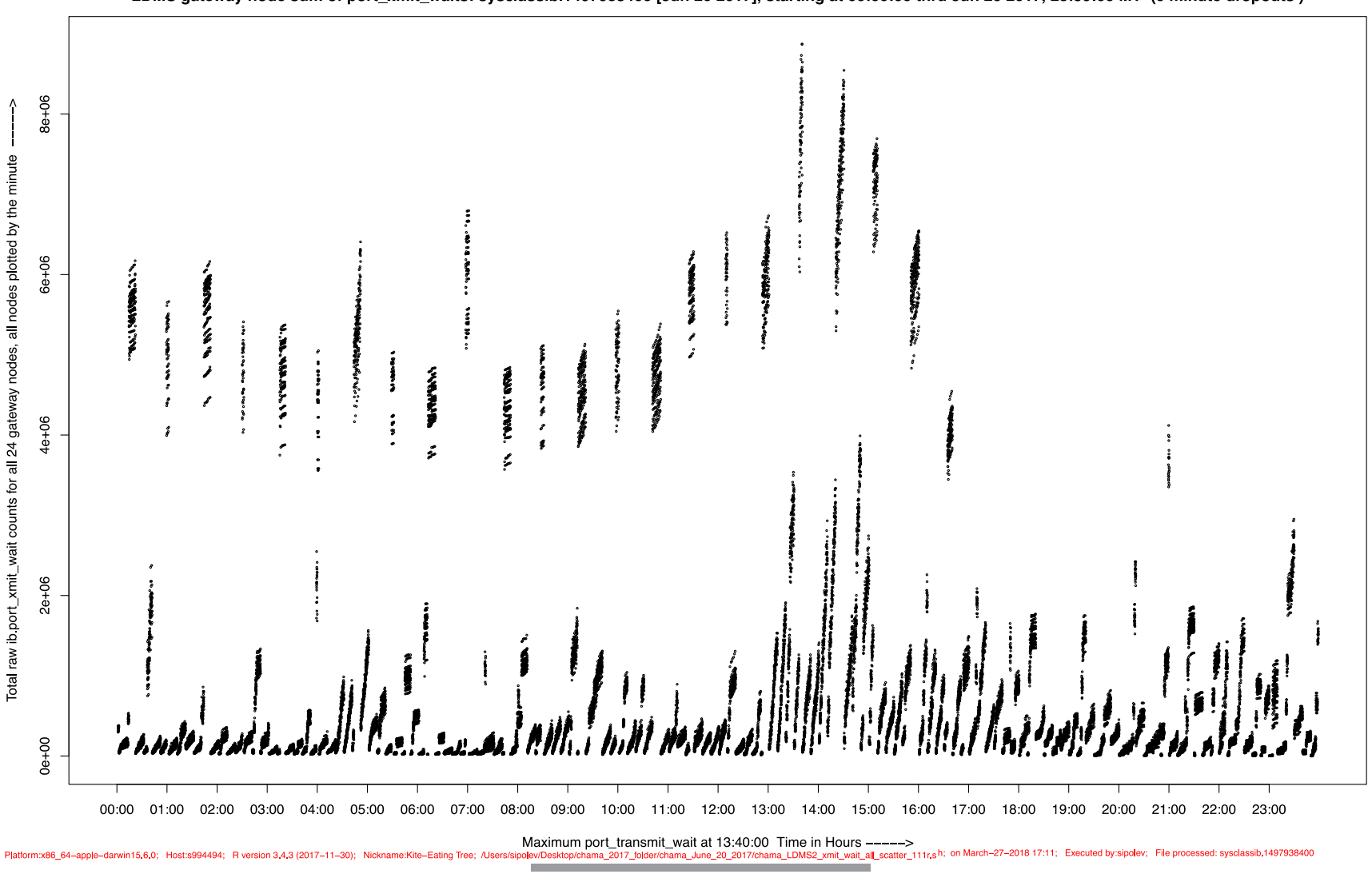
Switch

lustre® File System

IBM® GPFS

ALL GATEWAY PTW COUNTS

Exhibit H



VERIFYING PTW COUNT

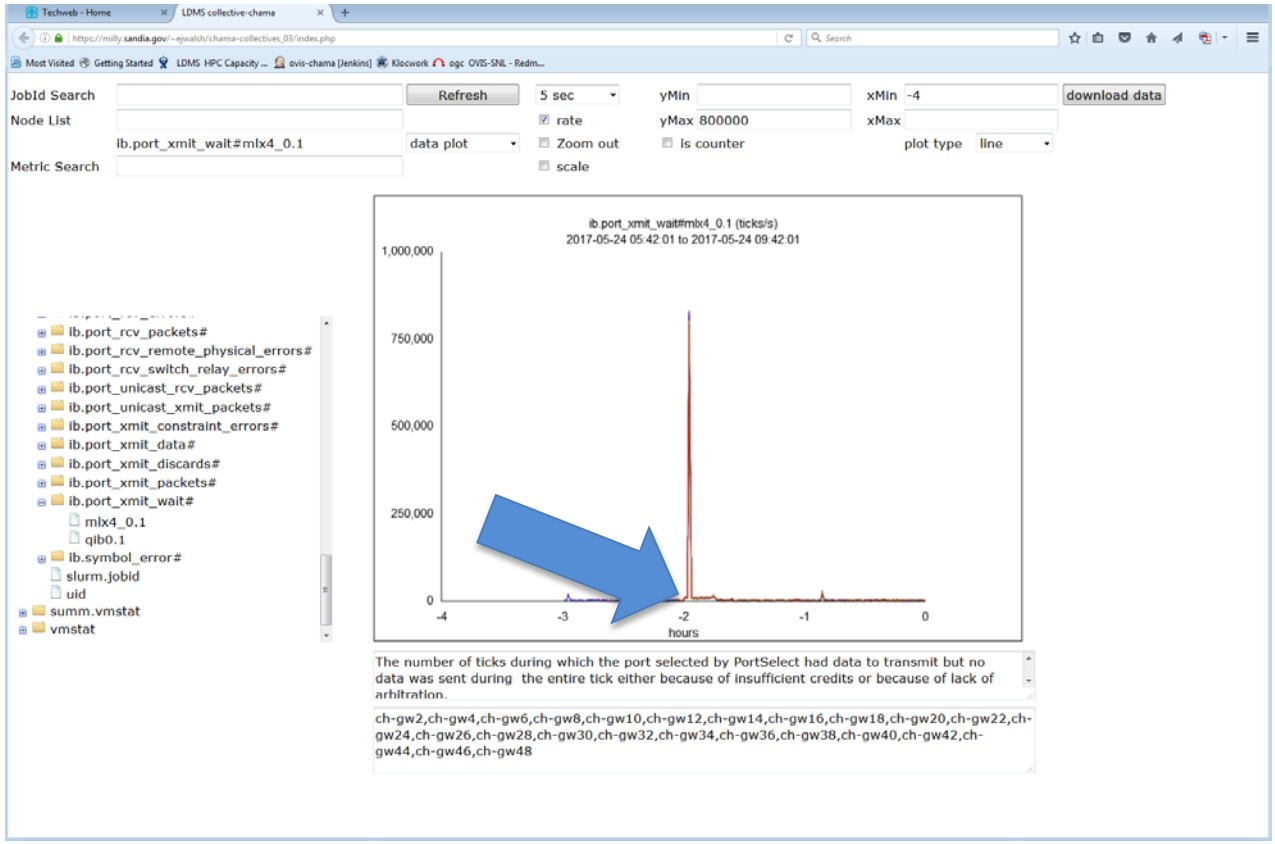
- **How Can We Verify Any of PTW Data ?**
- **We Do Need to Validate [“Trust – But Verify”]**
- **We Should Have 2nd Source [Ben Bradlee’s* Mantra] ?**
  No -- But We Do Have “Indirect Verification” -- *i.e.*,
  from a *crontab* Entry Running 2 Hours @ 5 mins Past

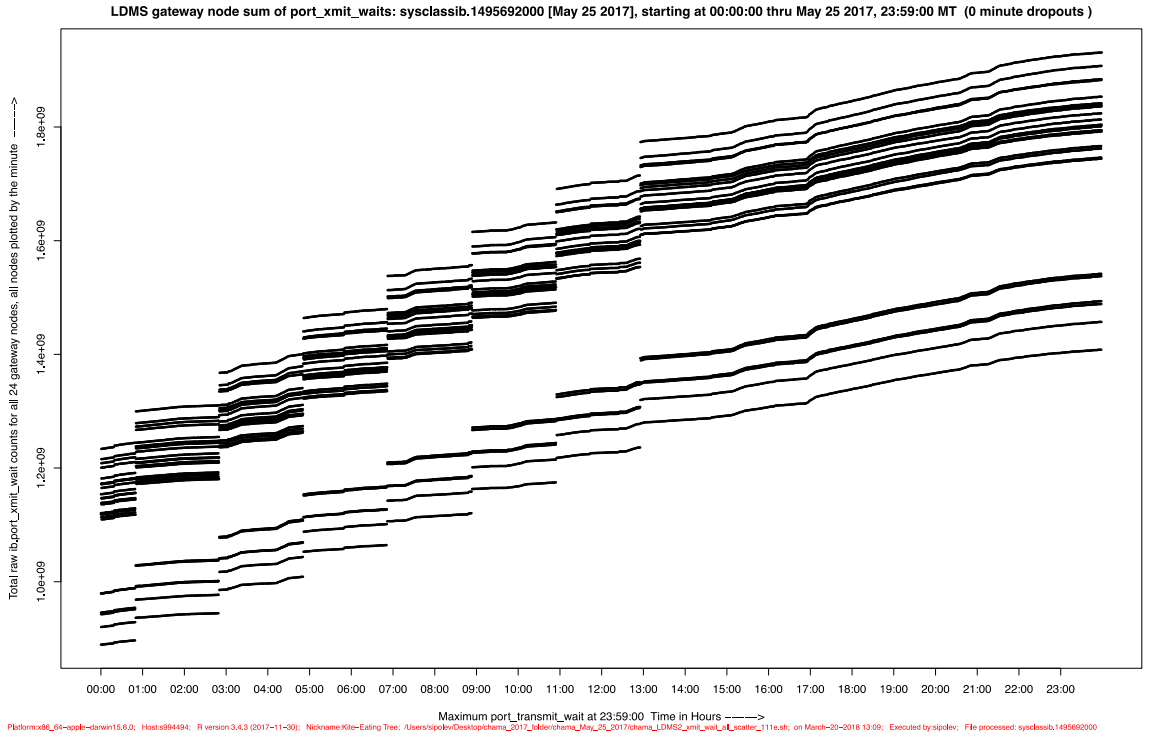
```
5 */2 * * * root   /..... /ib-edge-monitor
5 */2 * * * root   /..... /ib-core-monitor
```

INDIRECT PORT TRANSMIT WAIT VERIFICATION

- **Visually(and Also Confirmed from Looking at the Data)**

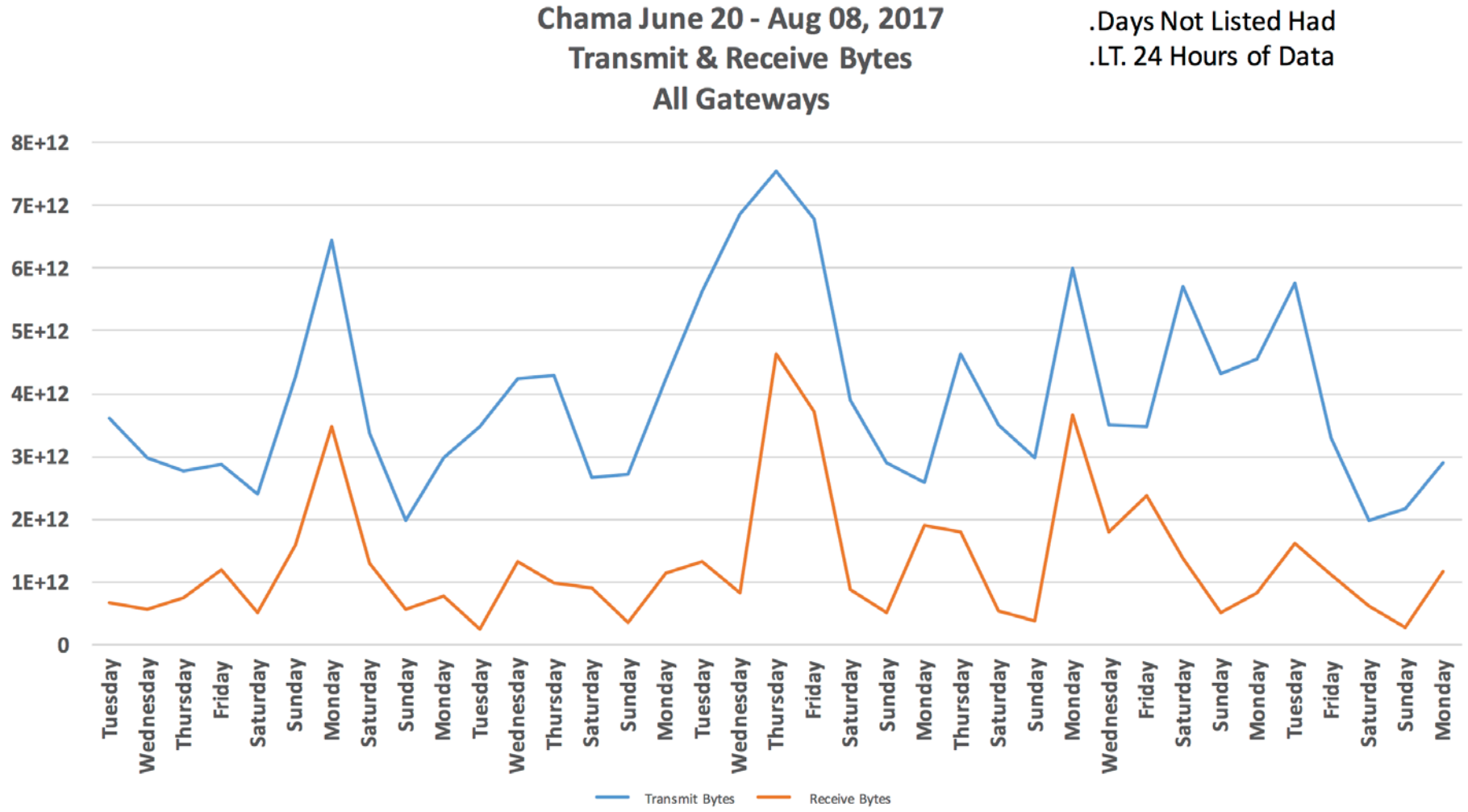
Exhibit [E]:





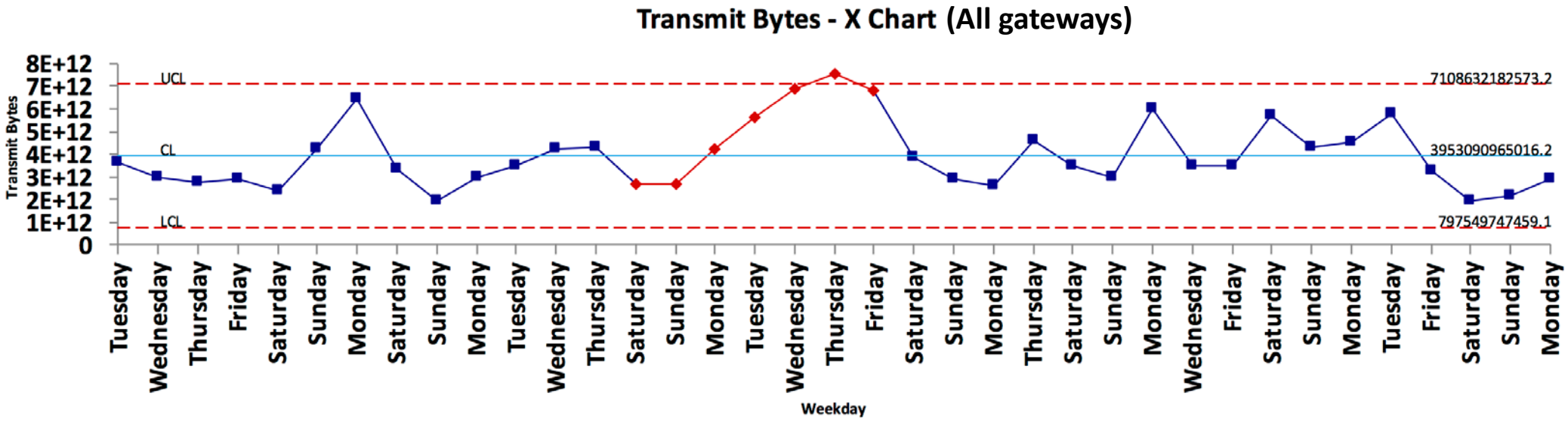
RUN CHART —EXPLORING OTHER FABRIC DATA

Exhibit I



SPC CHART EXPLORING FABRIC DATA

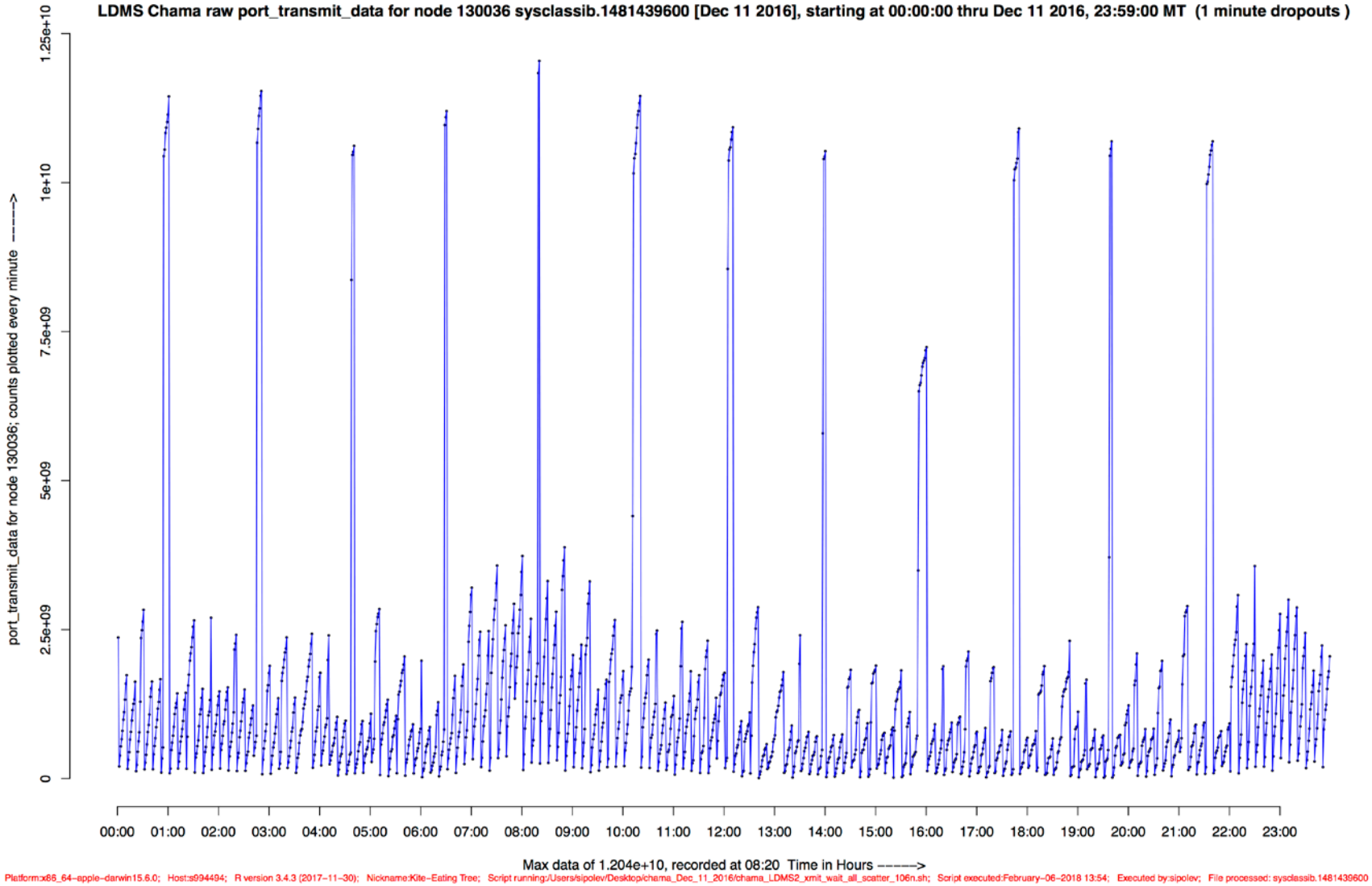
Exhibit J



- UCL (Upper Control Limit) & LCL (Lower Control Limit) Are 3 SD from the Average or CL (Center Line)
- Values Outside UCL & LCL Indicate “Special Cause” Is Impacting
- Values Inside —> “Common Cause”
- Red Points Indicate a Trend, Something to Investigate

PERIODICITY CHART — TRANSMIT DATA

Exhibit K



COMBINING THE PREVIOUS THREE CHARTS

- **The “Run” Chart (Showing Both Write & Reads) Gives Overview**
- **Next, the SPC Chart Shows If You Can Reasonably Predict (or Not) Future Behavior … But (Like Run) Also Loses Important Information**
- **Periodicity Chart Reveals “Pulses & Spikes” [Good to Know]**
- **All Three Needed to Help Predict/Manage/Design-for-Future**
- **Similarly, Both LDMS & “other tools” Needed to Cross-Check**

POSITING: WHAT WE *MAY* HAVE LEARNED

✓ Assumption: Gateways See Interesting Traffic

- ❑ Gateways Seem to Be Prone to PTW HOL Blocking {PTW Counts}
- ❑ No HCA Port_Transmit_Discards (Haven't Looked Everywhere)
  PTDs Have Been Seen in Switches
- ❑ Odds Are(?): HCA-to-HCA Writes Are Source of HOL Blocking*
- ❑ Certainly Need More Investigation & LDMS Will Be a Great Tool

✓ To Reduce Gateway PTW Counts & VL15 Drops

- ❑ Add More VLs (Add More Queues & Memory – But Complicated !)
- ❑ Just Add More Memory for Both Gateway HCAs:
  Not for More Connections, but More Memory for Each Connection

IN SUMMARY

- ✓ **We Have Lots of Data**
  - ❑ **LDMS Sees "Obvious" Errors (Symbol Errors, Link Down, *et al*.)**
  - ❑ **We Are Starting to See "Trends" and Also Pulses & Spikes**
  - ❑ **There Are Now "Subtle Values" that We Can Watch, & We Need INSIGHT as to Which "Subtle Values" (Port Transmit Wait, VL15s Dropped, *etc*.) Are Important & What Are "Tipping Points"**
- ✓ **LDMS v4 Is Coming … Adds OPA & Switch Support**
- ✓ **Hunt for "Which Values" & "Tipping Points" is On**
- ✓ **Come Join the Hunt !**

OPENFABRICS ALLIANCE

14th ANNUAL WORKSHOP 2018

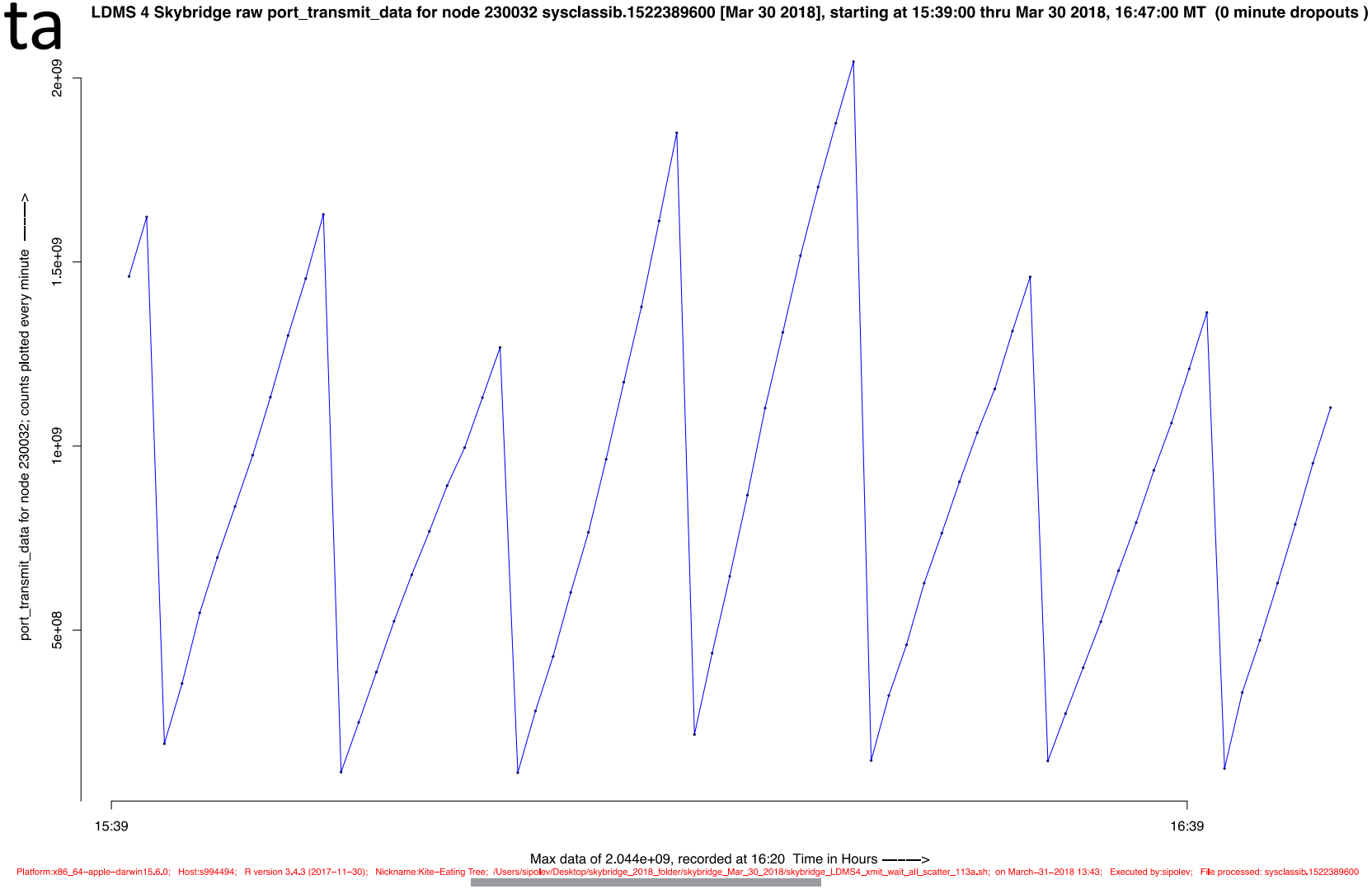
THANK YOU ! ... ANY QUESTIONS / DISCUSSION ?

Again, Thanks to Mike Aguilar, Ben Allan, Char Arias, Jay Livesay, Justin Wood, and Many Others !

Sandia National Laboratories

LAST MINUTE ADDITION #1

LDMS v4 Data



LAST MINUTE ADDITION #2

LDMS v4 Data