NINETEENTH ANNUAL

BiTS

Burn-in & Test Strategies Workshop

March 4 - 7, 2018

Hilton Phoenix / Mesa Hotel
Mesa, Arizona

Archive

COPYRIGHT NOTICE

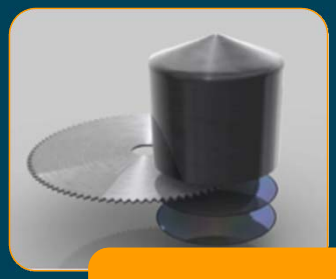
www.bitsworkshop.org

At Scale Cluster Validation

Antonio Villa & Victor Rodriguez Bahena
Intel

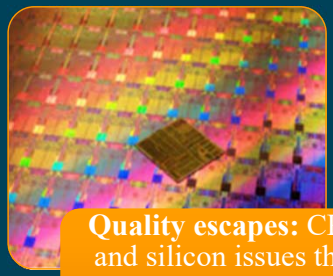
BiTS Workshop
March 4 - 7, 2018

Background

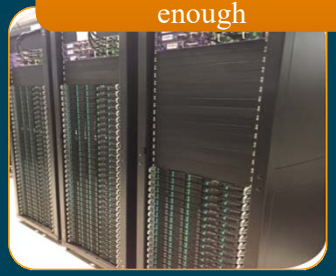


Design and silicon process complexities are growing

Intel DPM (Defects per million units) numbers: typical CPU goal is 500-1000 DPM



Quality escapes: CPU and silicon issues that Intel customers at the server industry that build datacenters are experiencing



Traditional validation methods need to evolve to meet advertised DPM numbers. Validation methods have not scaled enough

BiTS 2018 Burn-in & Test Strategies Workshop

At Scale Cluster Validation

2

Problem Statement and Goal

Intel server customers are experiencing **quality escapes** in large-scale server installations as a result of:

- "High Mean Time Between Failures" functional bugs
  - Increasing probability of hitting bug as CPU count increases
- Circuit marginalities
  - Cluster failures due to circuit marginality
- Manufacturing defects
  - Test hole in manufacturing test program

NOT EASY TO FIND

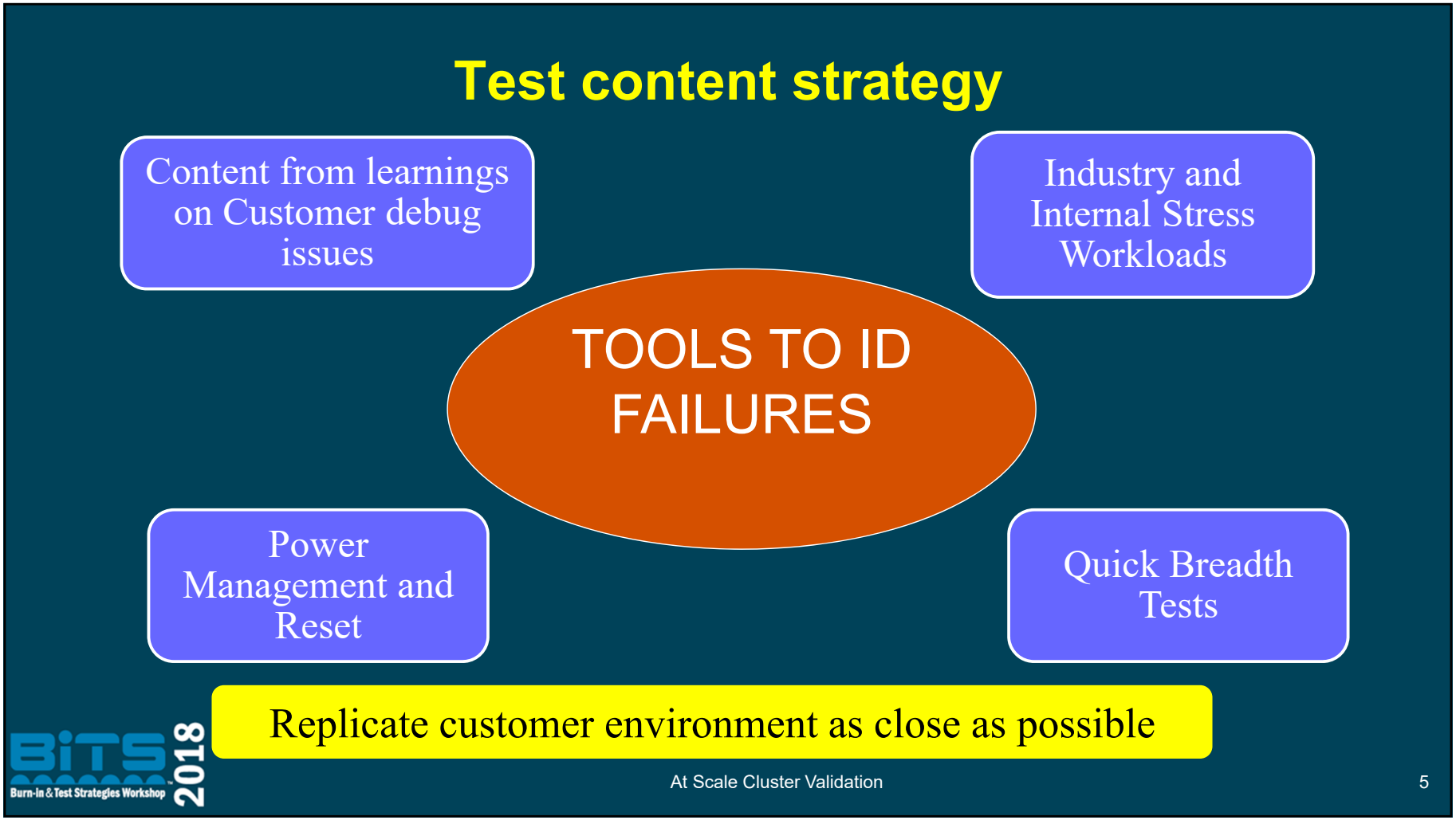
Need to find these issues before our customers!

Innovation Description

- Need to scale validation to a similar environment of the real customer datacenter solutions
- Need to find issues automatically in large number of systems
- Implement Telemetry solution:
  - automation of failure logs
  - CPU performance counters
  - Remote access
  - Comparison between nodes

The telemetry client is already released as open source by SSG OTC Intel group with the aim of making it replicable, adaptable and scalable at an economical cost for other teams and OEMs to satisfy their current needs
https://github.com/clearlinux/telemetrics-client
https://github.com/clearlinux/telemetrics-backend

(Right parts + right content) @ scale with automated telemetry


Test content strategy
Content from learnings on Customer debug issues
Industry and Internal Stress Workloads
TOOLS TO ID FAILURES
Power Management and Reset
Quick Breadth Tests
Replicate customer environment as close as possible
BiTS 2018 Burn-in & Test Strategies Workshop
At Scale Cluster Validation
5

Solution (1/2)

A telemetry client is a communication process by which measurements and data are collected at remote points and transmitted for monitoring and analysis.

Data center customers are savvy in collecting data to assess the health and performance of their clusters.

This includes the ability to identify "outlier" nodes in a large cluster

With the intention of finding these issues before our customers it was decided to adapt the telemetry solutions from SSG Open Source technology center.

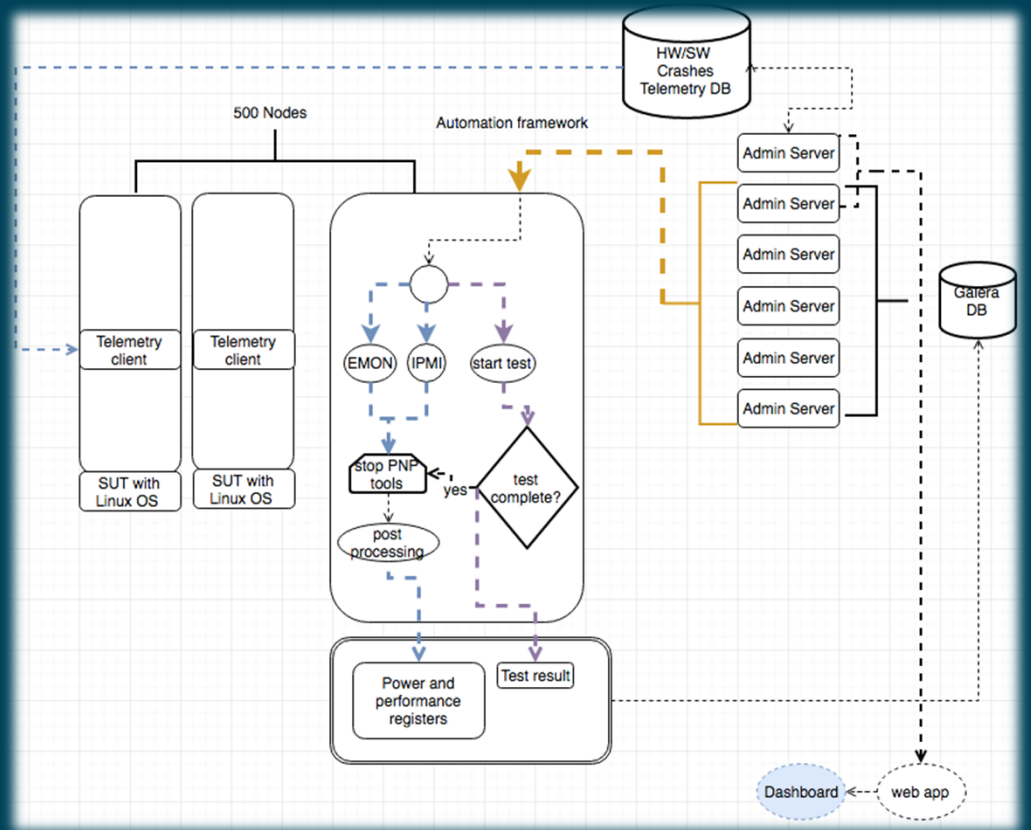
Solution (2/2)

This solution provides the component of a remote telemetry for Linux-based operating systems.

This telemetry system supports the detection and monitoring of software and hardware problems, from machine checks errors (MCE) to kernel crashes and application crashes

The way it collects logs and account machine checks errors on modern x86 Linux systems is through the MCELOG tool.

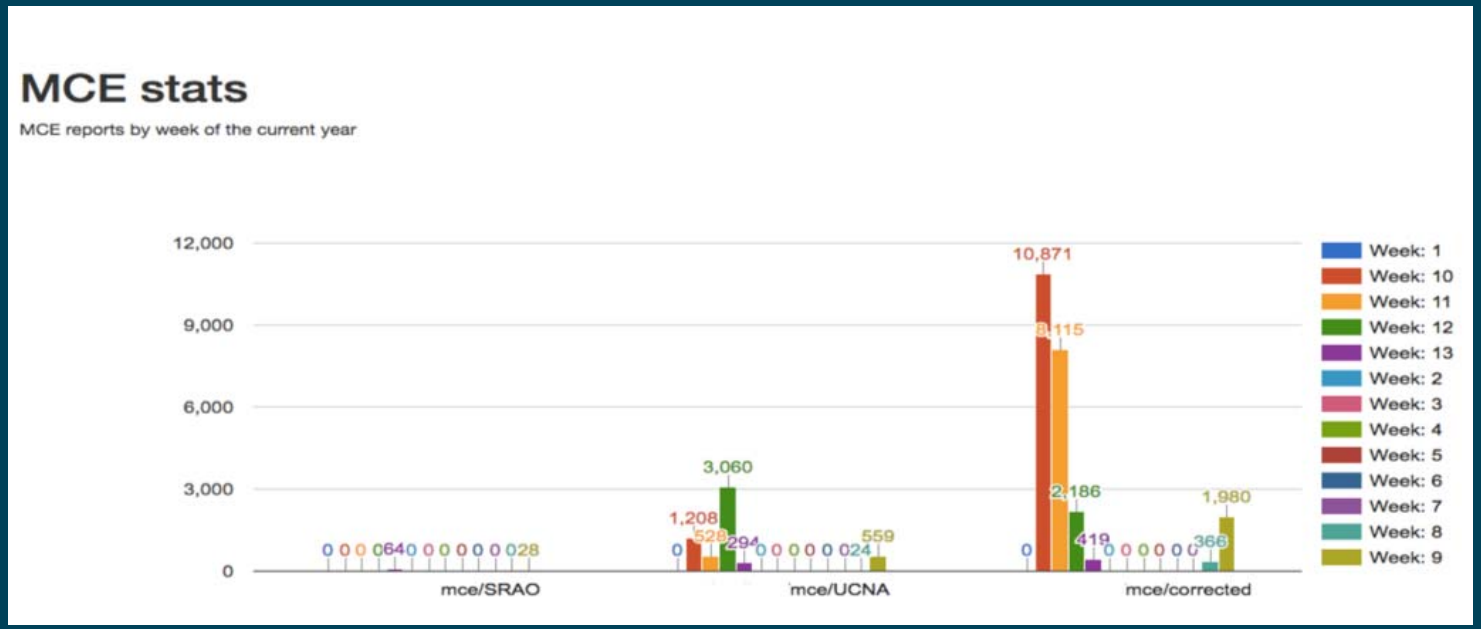
In order to catch CPU performance counters we use IPMI and EMON information during test execution



- Telemetry solution general architecture

Results/Potential (1/3)

- One of the immediate results that we have take advantage is the track of MCE over time as show below

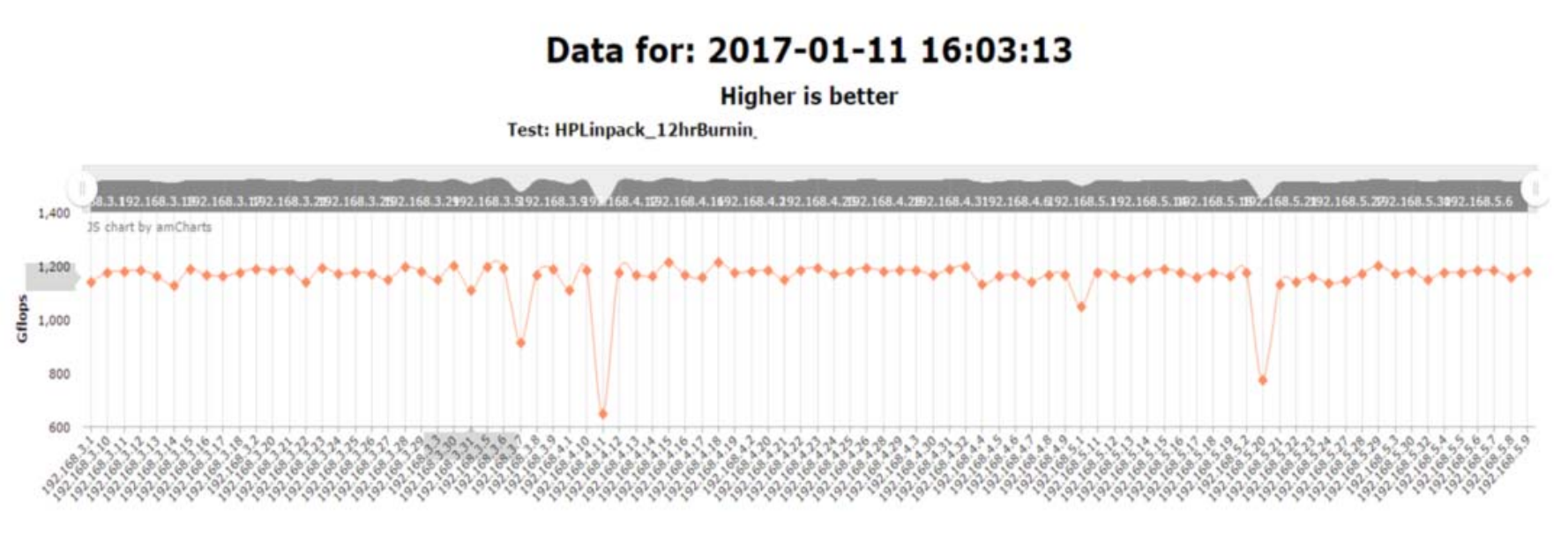


Machine Check Errors in cluster over time

At Scale Cluster Validation

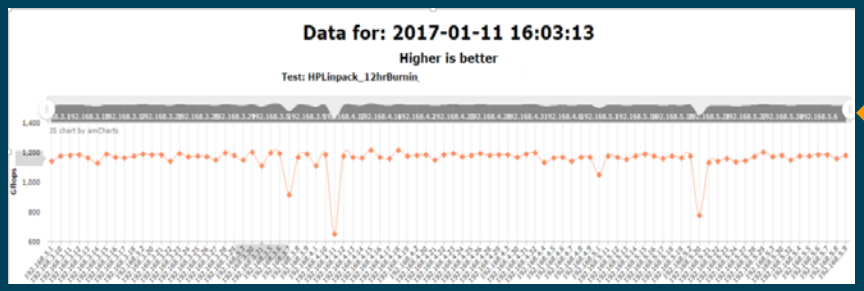
Results/Potential (2/3)

- We also used OTC tools to collect power and performance information which allows us to identify "outlier" nodes in the cluster just as problems reported by costumers in their data centers



Telemetry data showing three "outlier" nodes

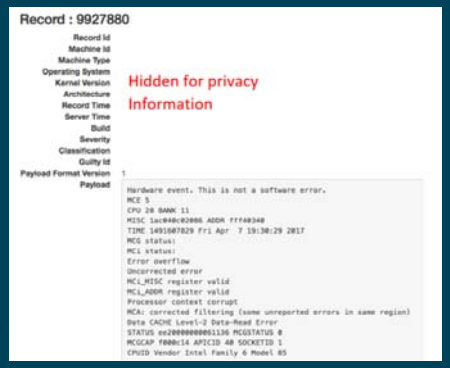
Results/Potential (3/3)



- 1.- Once an "outlier" node is identify after test execution

- 2.- Is easier to start the debug of the root cause analyzing the EMON and IMPI register captured during test execution

machine_clears_count_0	1989420.3248	2045382.1762	1611522.9296
qpi_0	99999999.9999	99999999.9999	99999999.9999
tsc_0	99999999.9999	99999999.9999	99999999.9999
cpu_freq_1	2.3145	2.3071	1.9495
cpu_utilization_1	99.0325	99.0337	99.1755
cpu_utilization_kernel_1	0.8239	0.7508	1.2296
cpi_kernel_1	5.3094	5.1719	6.6188
cpi_1	1.0495	1.0428	1.0313
emon_event_mux_reliability_1	99.8806	99.8639	97.9977
dram_power_1	17.5582	11.0004	11.8882
package_power_1	165.7072	168.1482	154.6837



- 3.- At the same is possible to find out if a MCE happened during the execution of that test in the "outlier" node
- 4.- Machine Learning and Big data algorithms are the next step in order to analyze all this data

Current Status

This Telemetry system is deployed at the 1000-CPU validation cluster at Guadalajara Intel Design Center site

Telemetry solution being used and prepared for future Intel products that will perform cluster validation

Telemetry system is helping to recreate costumer environments and being able to find these issues before them

It collects all necessary information required to automatically detect hardware issues remotely over the execution of the tests as well as CPU performance counters for debugging

Picture of Validation Racks with 500 nodes at GDC Intel Site

Acknowledgments

- Main Developers of telemetry solution: Victor Rodriguez, Gabriel Briones
- Collaborators on documentation and development: Robert Nesius , Patrick McCarty, Mathew Johnson, Victor Rodriguez
- Cluster validation team: Jan Glott, Jim Rowan, Miguel Figueroa, Alberto Arechiga, Vikram N Chowdiah,