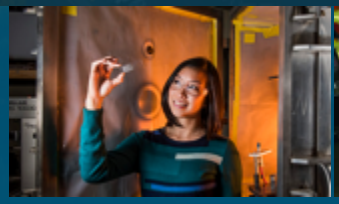
SAND2022-2446C

Sandia National Laboratories

Characterizing Failures in HPC Using Benford's Law



SIAM-PP 2022: Toward Scalable Resilient and Fault Tolerant Applications for Extreme Scale Computing Systems

PRESENTED BY

Kurt B. Ferreira & Scott Levy

CCR Center for Computing Research

DOE 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.

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.

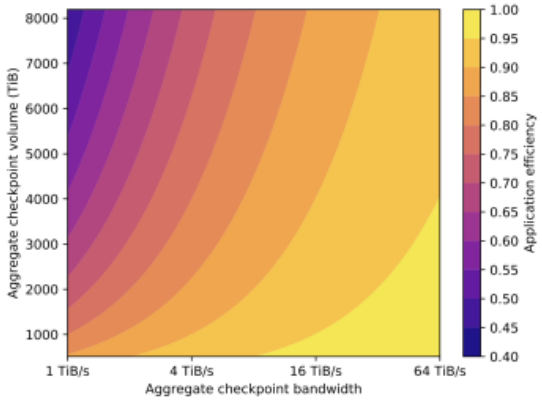
Key Take-Away

- Examining the lifetime of failures in the Cielo supercomputer that was located at the Los Alamos National Labs (LANL) and the Astra supercomputer at Sandia National Labs, we show that the time between faults on these systems obey Benford's Law (the leading digit is likely to be small).

- Exploiting Benford's Law may be useful in verifying, modeling and mitigating failures on HPC systems

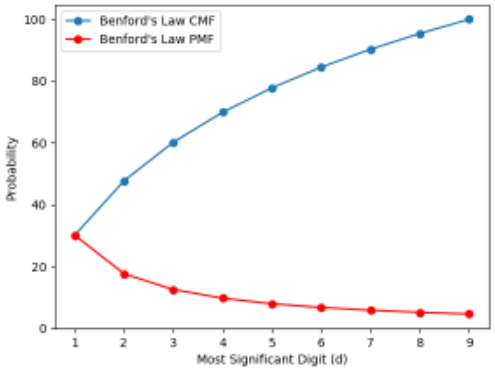
Characterizing Failures is Important in HPC!

- Characterization has lead to significant insights:
  - Optimal checkpointing intervals $P = \sqrt{2\mu_f C}$
  - Failure avoidance through prediction
  - Viability of checkpointing at exascale



What is a Newcomb-Benford Distribution?

- Also called the Newcomb–Benford law, the law of anomalous numbers, or the first-digit law, is an observation about the frequency distribution of leading digits in many real-life sets of numerical data. The law states that in many naturally occurring collections of numbers, the leading digit is likely to be small.

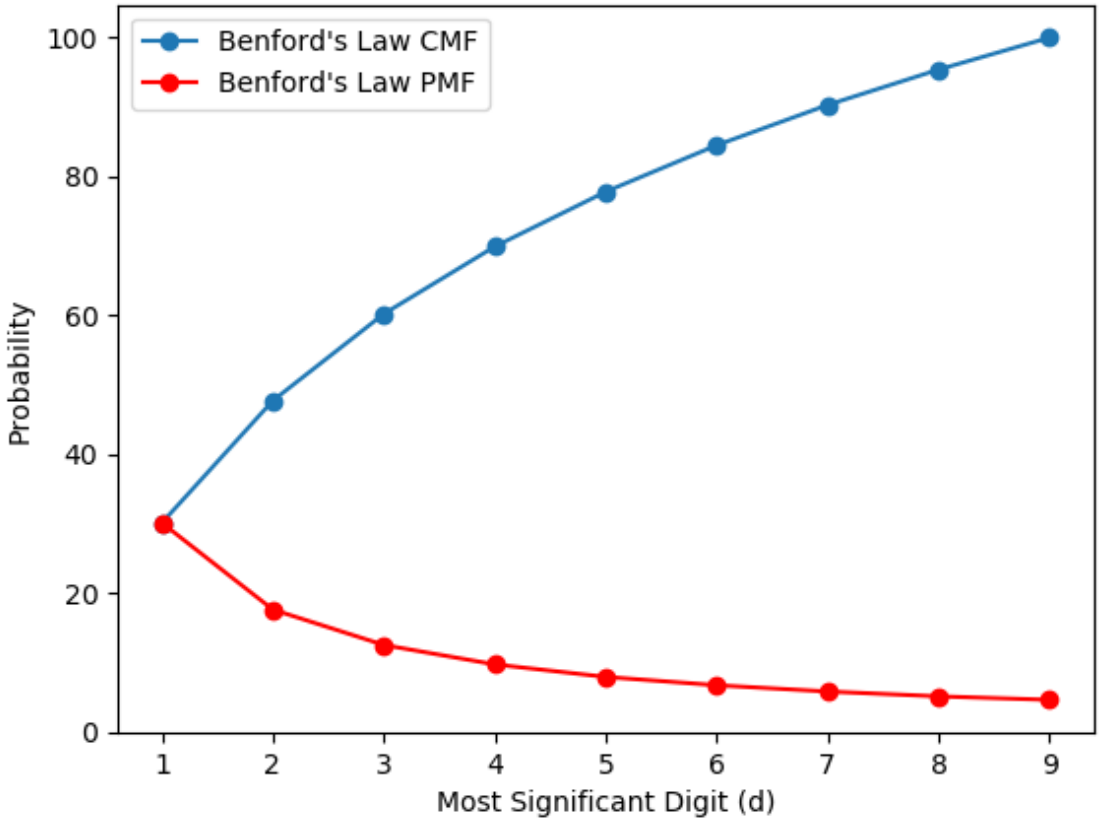


- Originally observed in 1881 by astronomer Simon Newcomb noticed that in logarithm tables the earlier pages (those that started with 1) were much more worn than the others.

- It has been shown that this result applies to a wide variety of data sets, including: electricity bills, street addresses, stock prices, house prices, population numbers, death rates, lengths of rivers, COVID infection rates, physical and mathematical constants, and is used for scientific fraud detection.

- Note: this property is independent of the time between failures of a node and is independent of representation (seconds, hours, etc.).

Newcomb-Benford Distribution CMF and PMF



Data Source: Memory Failures From Cielo (2011-2016)

Operational Life	March 2011 – May 2016
Location	Los Alamos, New Mexico, USA
Architecture	Cray XE6
Size	~8,500 compute nodes
Memory	32 GB/compute node
ECC	Chipkill-correct
Top500 rank	#6 (June 2011) -- #75 (June 2016)

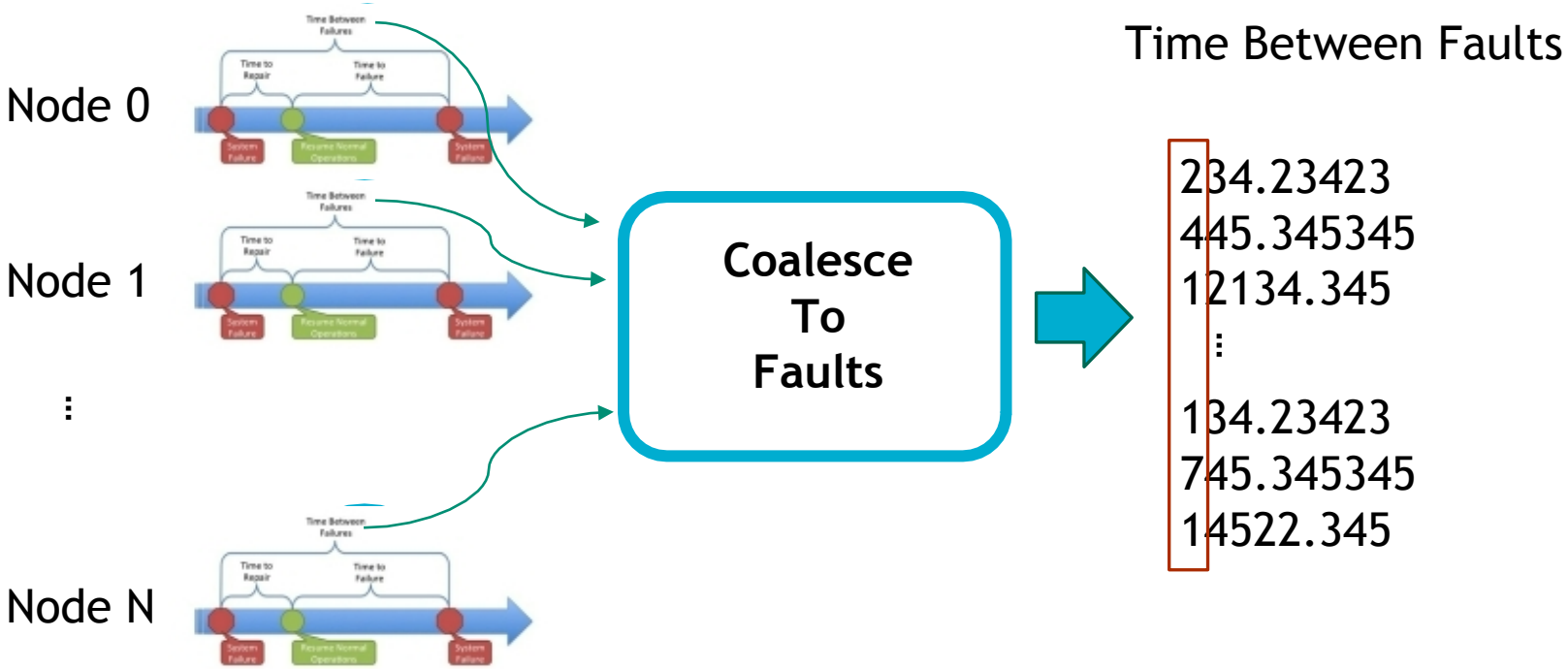
Data Source: Correctable Memory Failures From Astra

Dataset	Jan. 20, 2019 - Sept. 14, 2019
Location	Albuquerque, New Mexico, USA
Architecture	ARM ThunderX2 processor
Size/Cores	2,592/145,152
Memory	128 GB/compute node
ECC	SEC-DED
Top500 rank	#198 (11/2019) -- #393(11/2021)

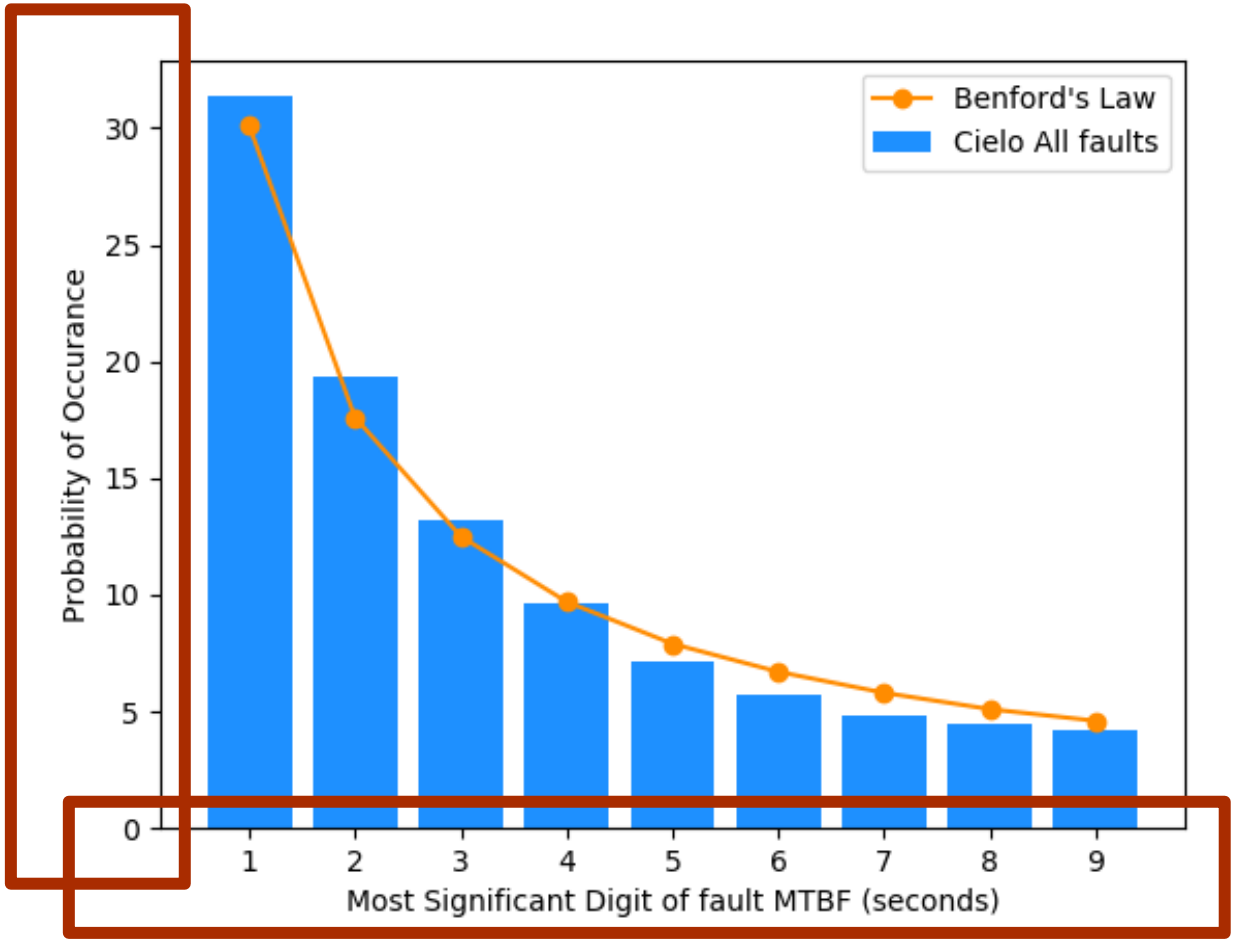
Terminology

- A fault is the underlying cause of an error (e.g., stuck-at 1 bits or high-energy particle strikes). Faults can be active (causing errors), or dormant (not causing errors).
- An error is incorrect system state due to an active fault. Errors are detected and possibly corrected by higher-level mechanisms such as parity or error correcting codes (ECC). They may also be uncorrected or, in the worst case, undetected (SDC).
- Transient faults cause incorrect data to be read from a memory location until the location is overwritten. These faults occur randomly and are not indicative of device damage (e.g. particle-induced upsets).
- Permanent faults cause incorrect data on each access.

Analysis: Collect the Most Significant Digits of the Time Between Faults

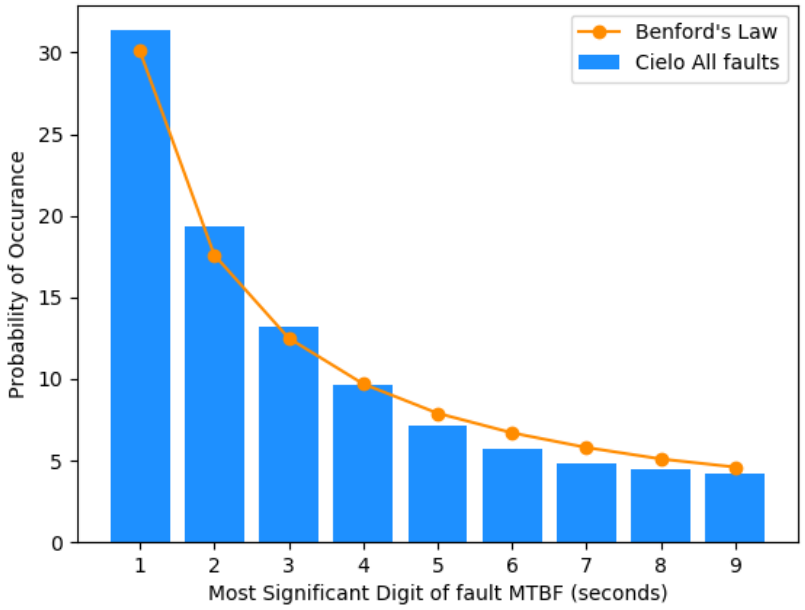


Interpreting Plots

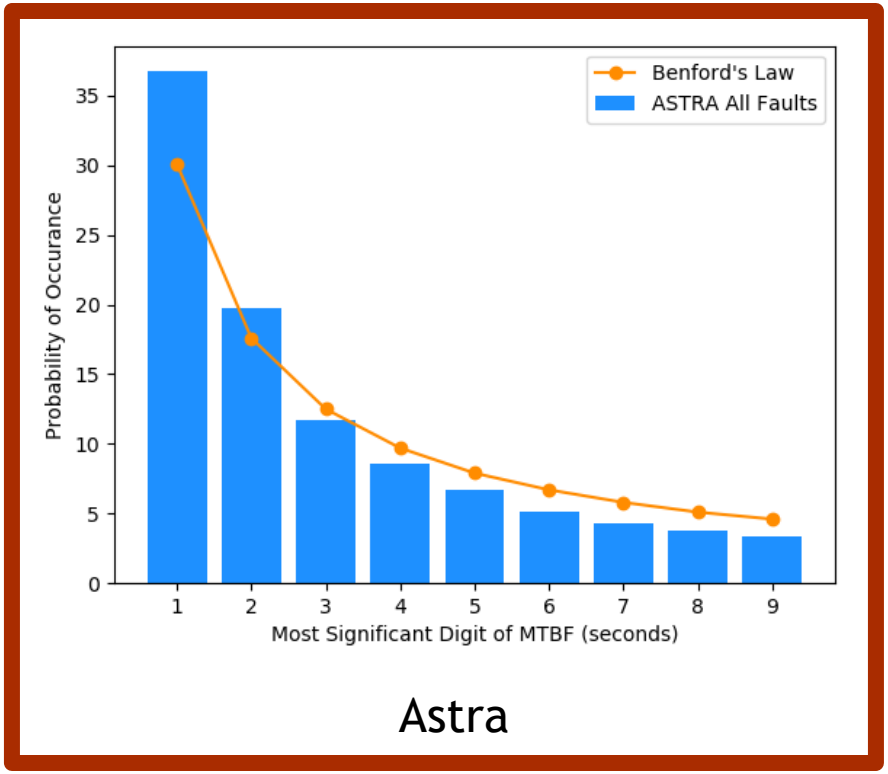


If MSD is zero (e.g. 0.14), use next lower order digit

Cielo and Astra: Faults follow Benford's Law



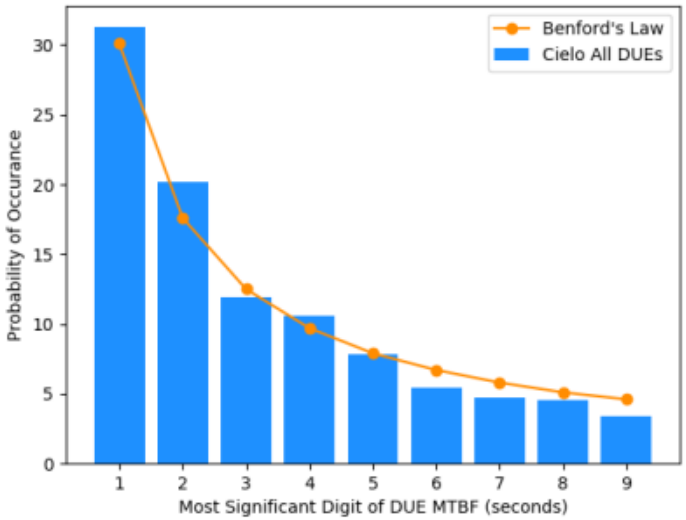
Cielo



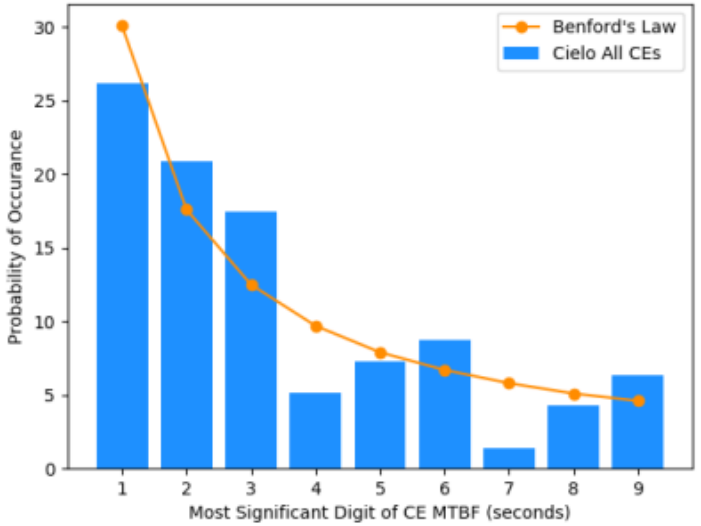
Astra

Variation from Benford: Only nine months of data as opposed to ~5 years for Cielo

Cielo: Uncorrectables Obey Benford's Law



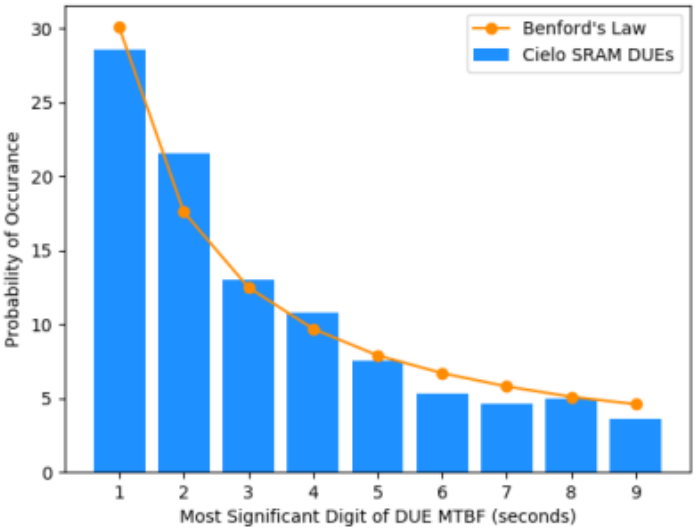
All Uncorrectable Faults



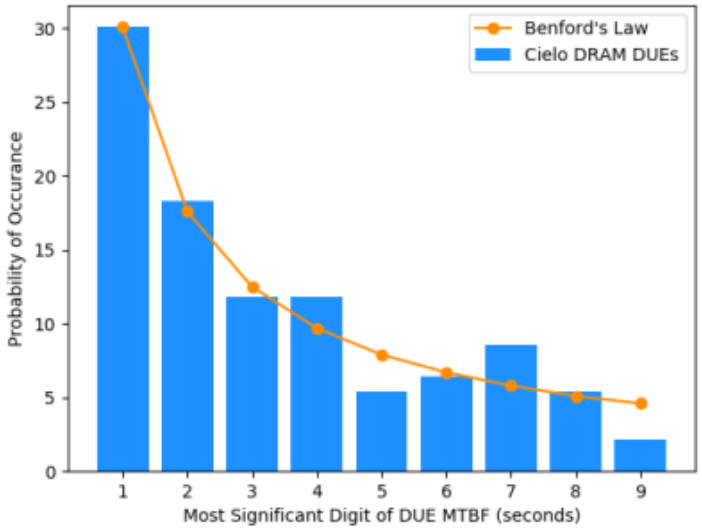
All Correctable Faults

Possible Explanation: correctable errors on Cielo were logged in a polled ring buffer therefore errors can be lost, Impacting the calculation of the fault time.

Cielo UC: SRAM Faults Obey Benford's Law



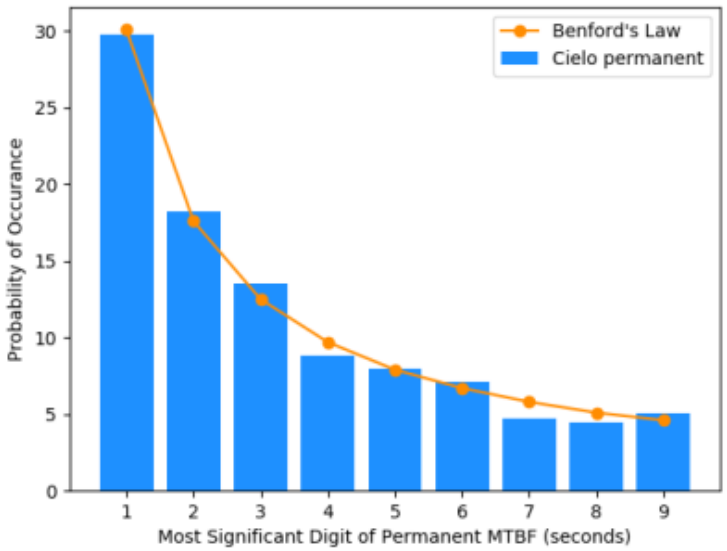
SRAM Uncorrectable Faults



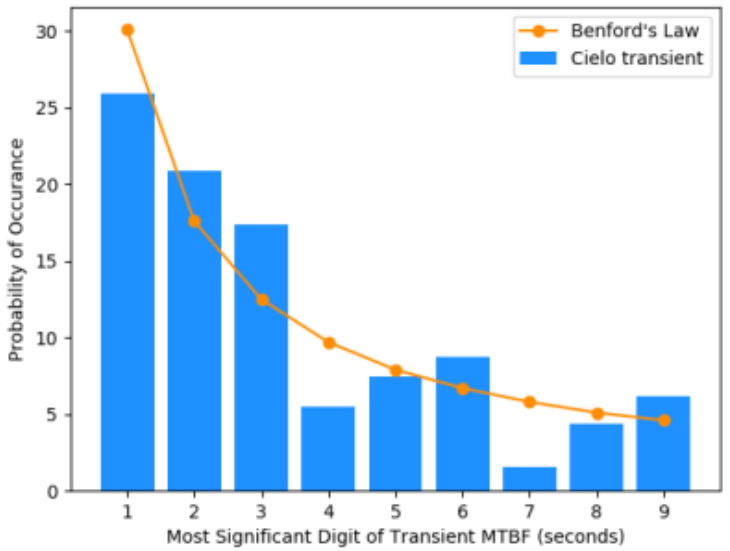
DRAM Uncorrectable Faults

Possible Explanation: few DRAM DUE in Cielo's lifetime, significantly more SRAM errors.

Cielo: Permanent Faults Obey Benford's Law



Permanent Faults



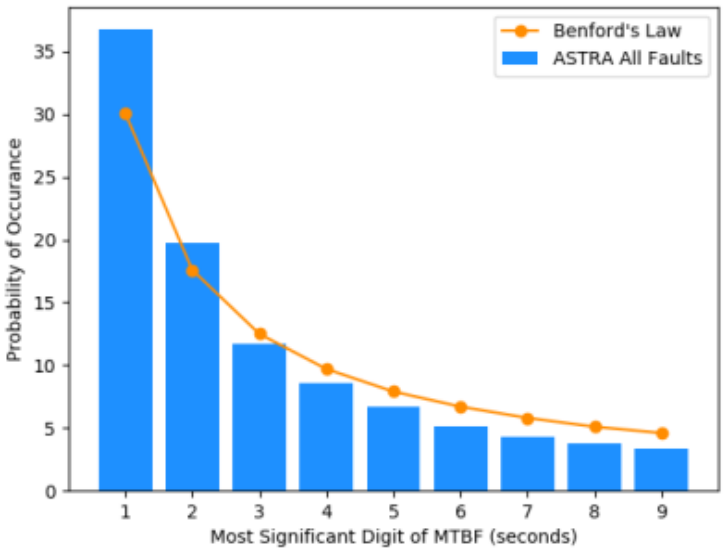
Transient Faults

Possible Explanation: the majority of permanent faults are uncorrectable and transient are correctable (polling impacts timing).

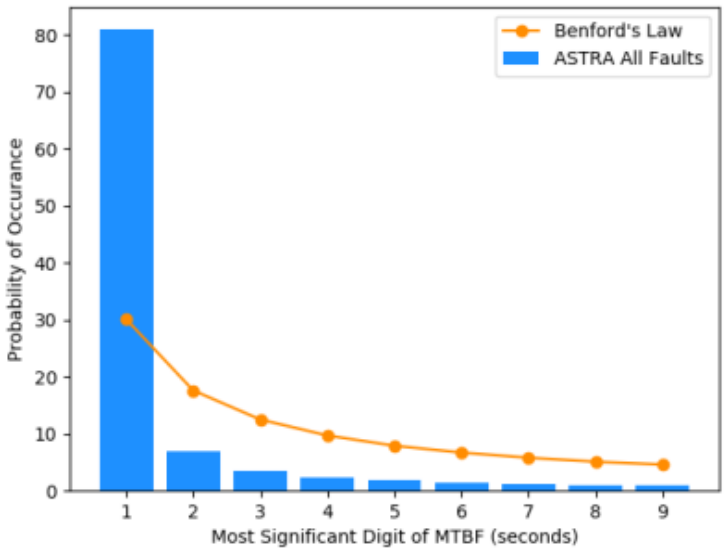
How Can We Use and/or Leverage Benford's Law?

- Verifying faults versus errors in failure data
- In modeling, choosing an appropriate random variable distribution
- In failure mitigation, as a cheap form of prediction.

Fault Distributions Appear to Follow Benford's Law



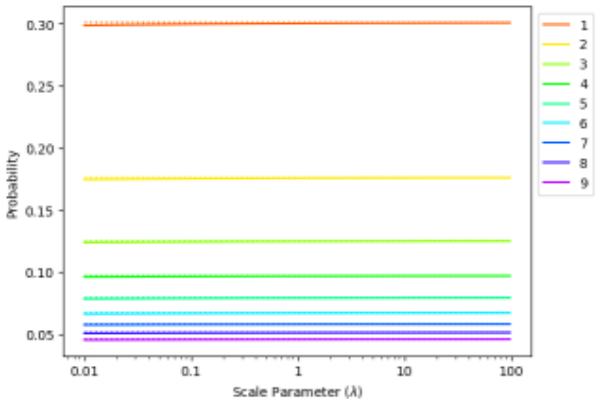
Faults



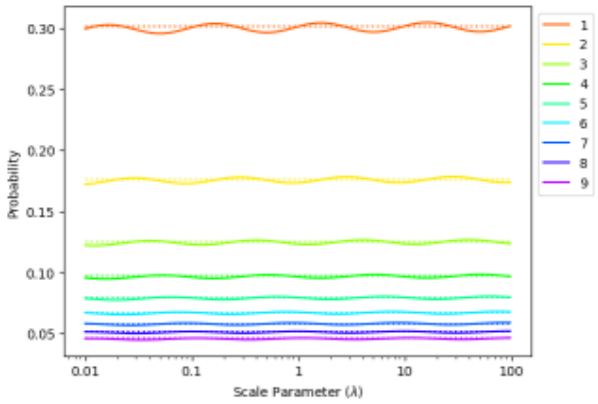
Errors

Errors are an artifact of an applications memory access pattern, not a platforms reliability.

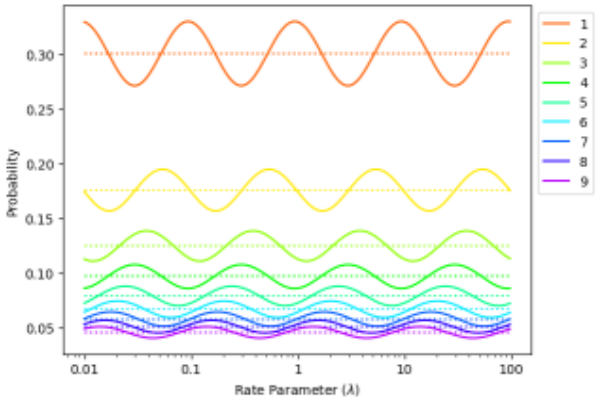
Some Distributions Follow Benford's Law Better Than Others



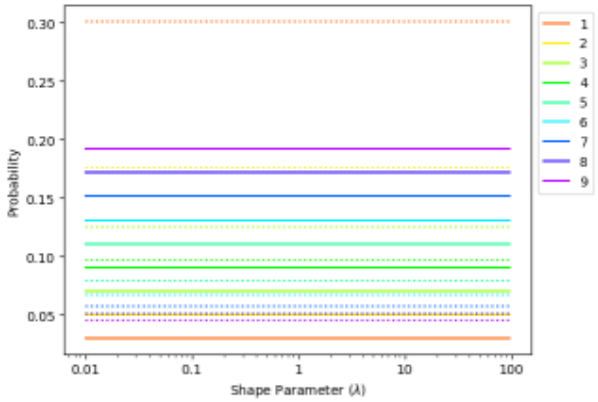
Weibull



Gamma

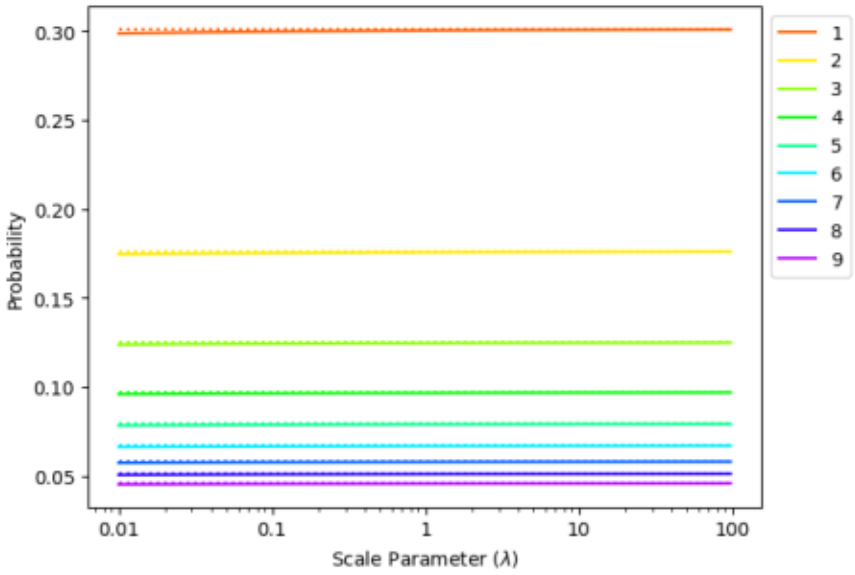


Exponential

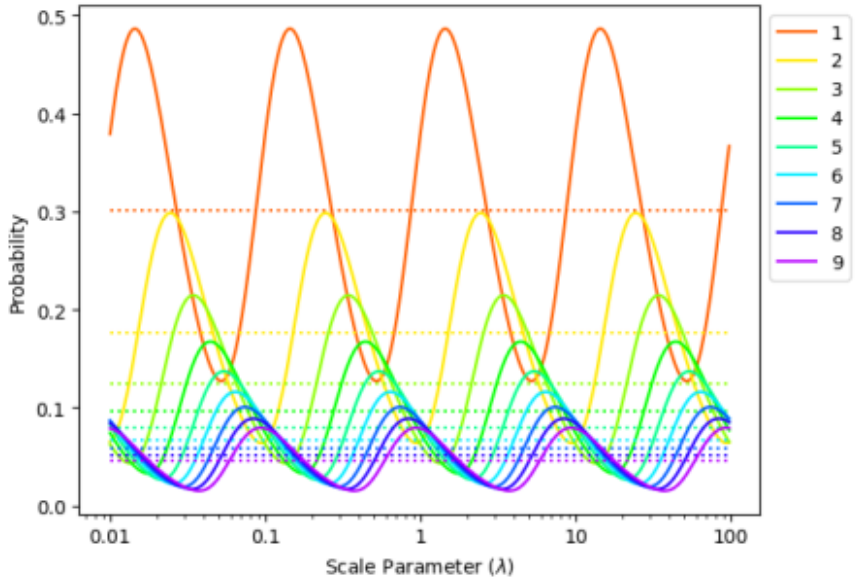


Power Law

Distribution Parameters Can Impact Distribution of MSD



Weibull, Scale = 0.25



Weibull, Scale = 2.0

Using Benford's Law for Failure Prediction?

- Very low overhead form of prediction, only need to observe the MSD of the MTBF
- Intervals where failures are more likely so adjust the checkpoint times accordingly
- This work is on-going, stay tunned!

Summary

- Failures from the Cielo and Astra supercomputers accurately follow Benford's law.
- Uncorrectable faults appear to obey Benford's law more closely than correctables due to the way these errors are logged on the system.
- Both SRAM and DRAM faults appear to obey Benford's Law, SRAM more closely due to significantly larger number of faults in sample.
- Benford's Law analysis of failure data for HPC may be useful in verifying fault analysis, choosing a more accurate random variable distributions in modeling, and as a simple form of prediction.

Kurt B. Ferreira
kbferre@sandia.gov

Questions?/Comments?

Scott Levy
sllevy@sandia.gov

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. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.