ARS 2005

San Diego, California, USA

Track 1, Session 9

Simple Plots for Monitoring Field Reliability

David Trindade, Ph.D.
Swami Nathan, Ph.D.
Sun Microsystems Inc.

International Applied Reliability Symposium
ARSymposium.org

Sun microsystems

Introduction

- Reliability is a key concern of Sun customers
  - What is the reliability ?
  - What should be the reliability ?
- MTBF (Mean Time Between Failures) is the typical metric used for communicating reliability.
- MTBFs imply many assumptions and are prone to misinterpretation.

Introduction

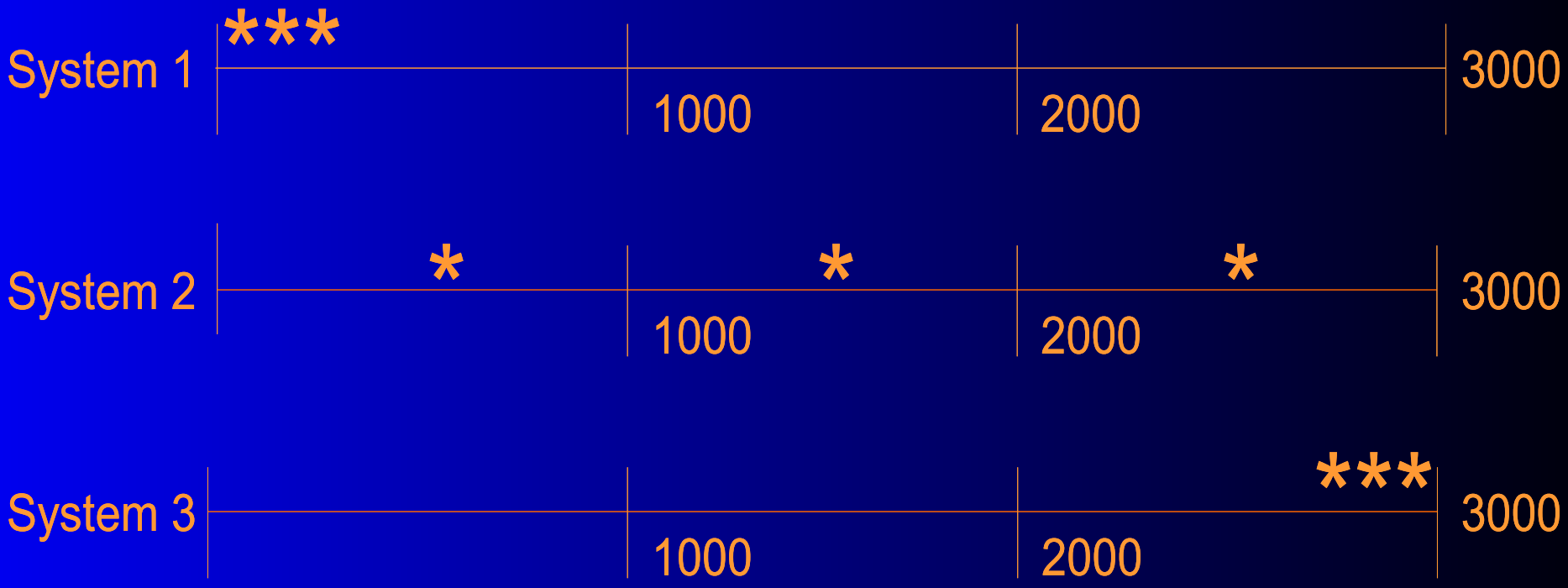
- Customers want to know more than MTBFs
  - What are the causes of downtime ?
  - What can we expect going forward ?
- Pareto, stacked bar, pie, and other static charts are often used to convey analysis results.
- Such charts can mislead by hiding important effects related to time.

Agenda

- MTBF Limitations
- Repairable Systems Analysis for Age and Calendar Time
- Time Dependent Cause Plotting
- Case Studies
- Summary

MTBF Hides Information

Example with 3 failures in 3000 hours...



MTBFs are the same, implying equal reliability.

Dangers of Extrapolating to an MTBF

- During the years 1996-1998, the average annual death rate in the US for children ages 5-14 was 20.8 per 100,000 resident population.
- The average failure rate is thus 0.02%/yr
- The MTBF is 4800 years!!

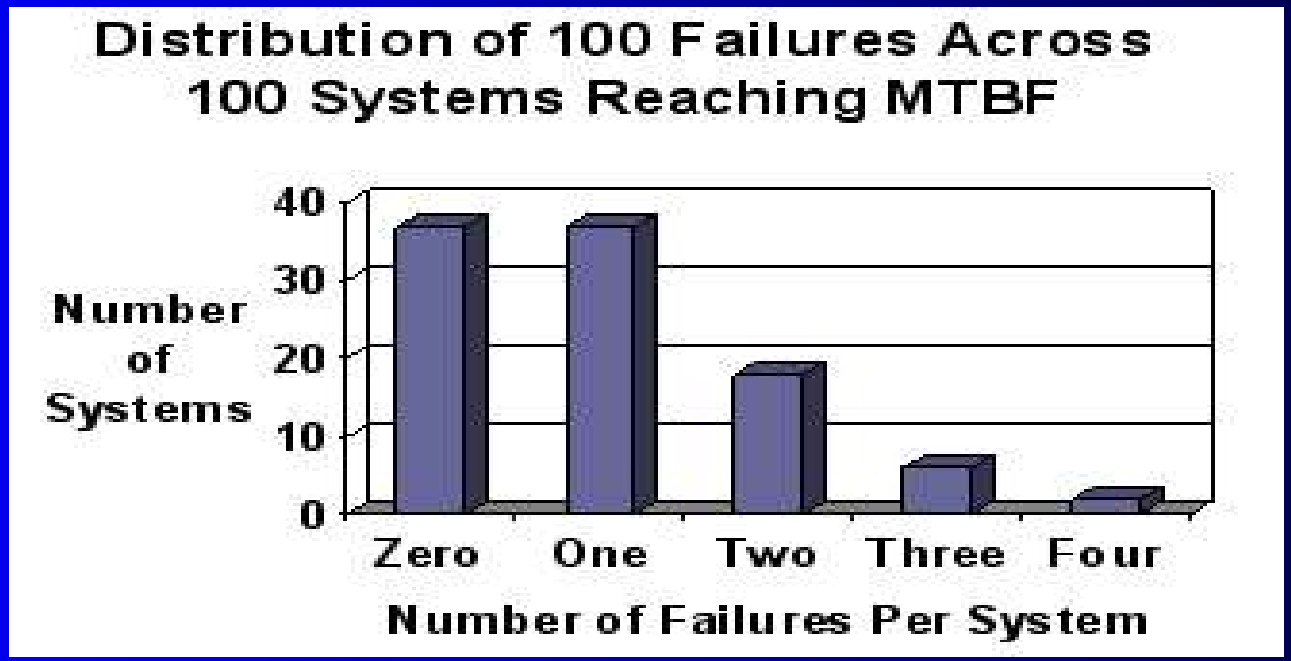
MTBF Assumptions: When is it OK ?

- Repairable Systems
- Renewal Process ("as good as new")
  - Times ***between failures are independently and identically distributed***
  - ***Single distribution of times between failures***
- ***Assume Exponential Distribution***
  - ***Constant hazard rate (time independent)***
  - ***No trend (constant recurrence rate RR or ROCOF)***
- ***Homogeneous Poisson Process***

MTBF Implications for HPP

For a 100 repairable systems, by the time they all reach the MTBF, on the average there will be 100 failures.

#Systems(# Fails/System)
37(0)
37(1)
18(2)
6(3)
2(4)



Customers can focus on the worst machines and perceive ***a reliability problem.***

MTBF-Inadequate Reliability Measure

- Valid only for a constant RR (HPP)
- Treats all system hours and all failures as equivalent and ignores age effects
- Data is rarely checked for validity of HPP

- Customers accustomed to MTBF usage
- MTBFs are often quoted with imprecise definition of failure (not = outage)
- We need a better and more accurate approach to measure reliability.

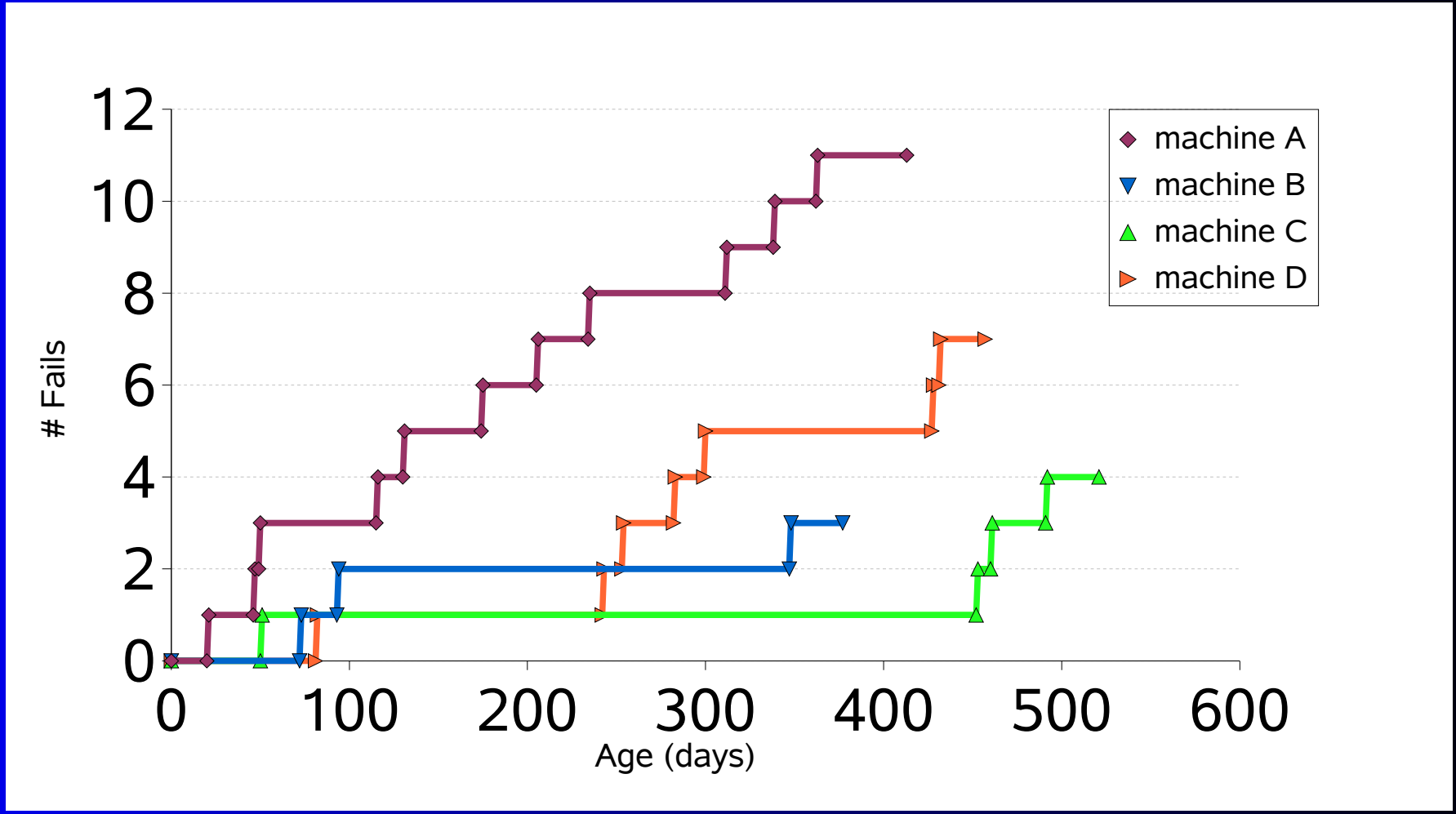
Parametric Methods

- MLE methods for HPP and NHPP are extremely powerful and rigorous.
- Lots of literature.
- Too complex for communicating with management and customers.
- Some customers tend to think "*information is hidden with statistical cleverness*".
- Showing a maximum likelihood equation is not the fastest way to gain credibility with customers.

Time Dependent Reliability (TDR)

- TDR is a Sun Microsystems acronym for non-parametric analysis of reliability data.
- Novice practitioners relate more easily to the non-parametric approach compared to indiscriminate modeling with various distributions.

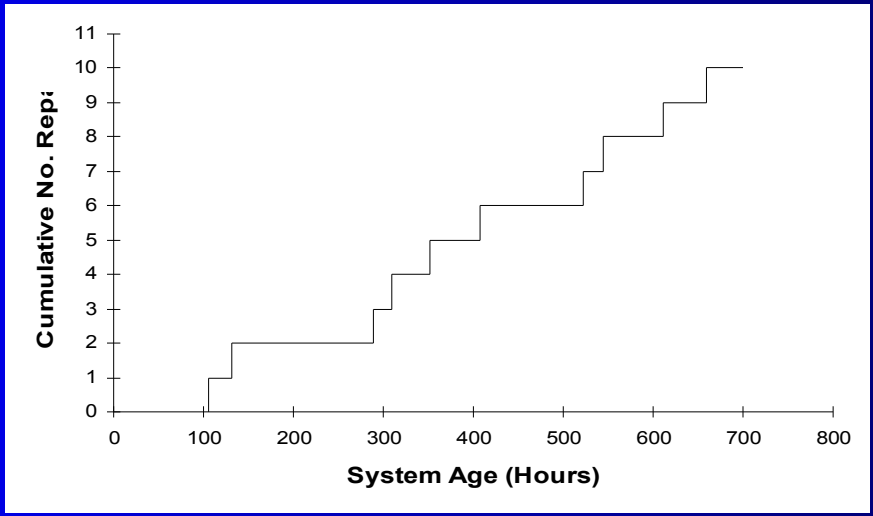
Cumulative Plot



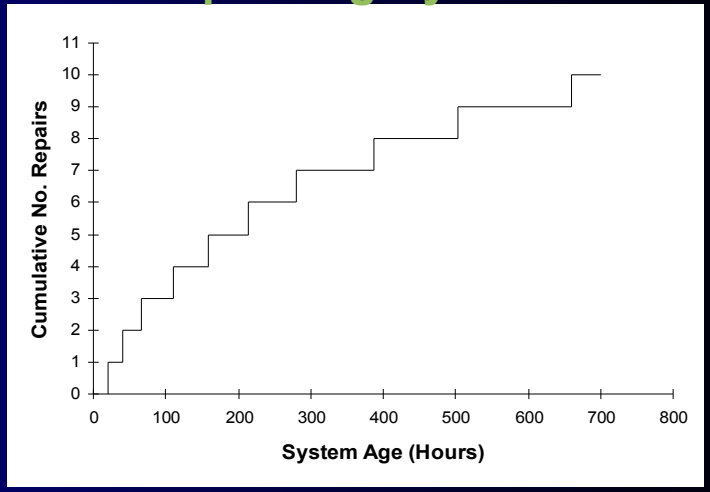
Cumulative plot shows failure history as number of failures (repairs) Vs time for each machine

Cumulative Plots reveal trends

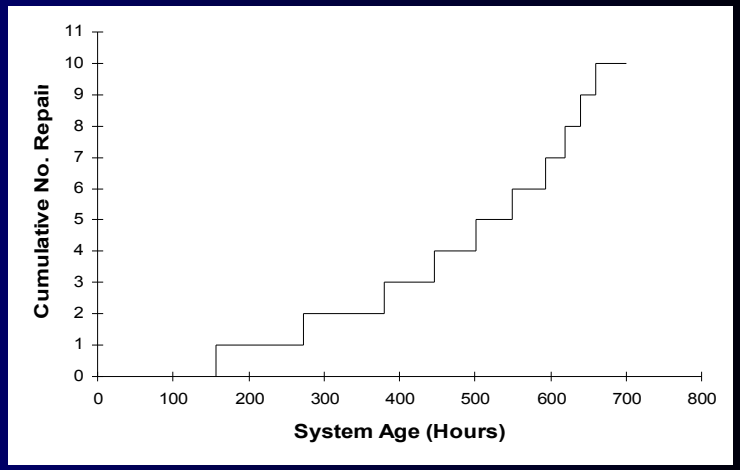
Stable System – No Trend



Improving System



Worsening System

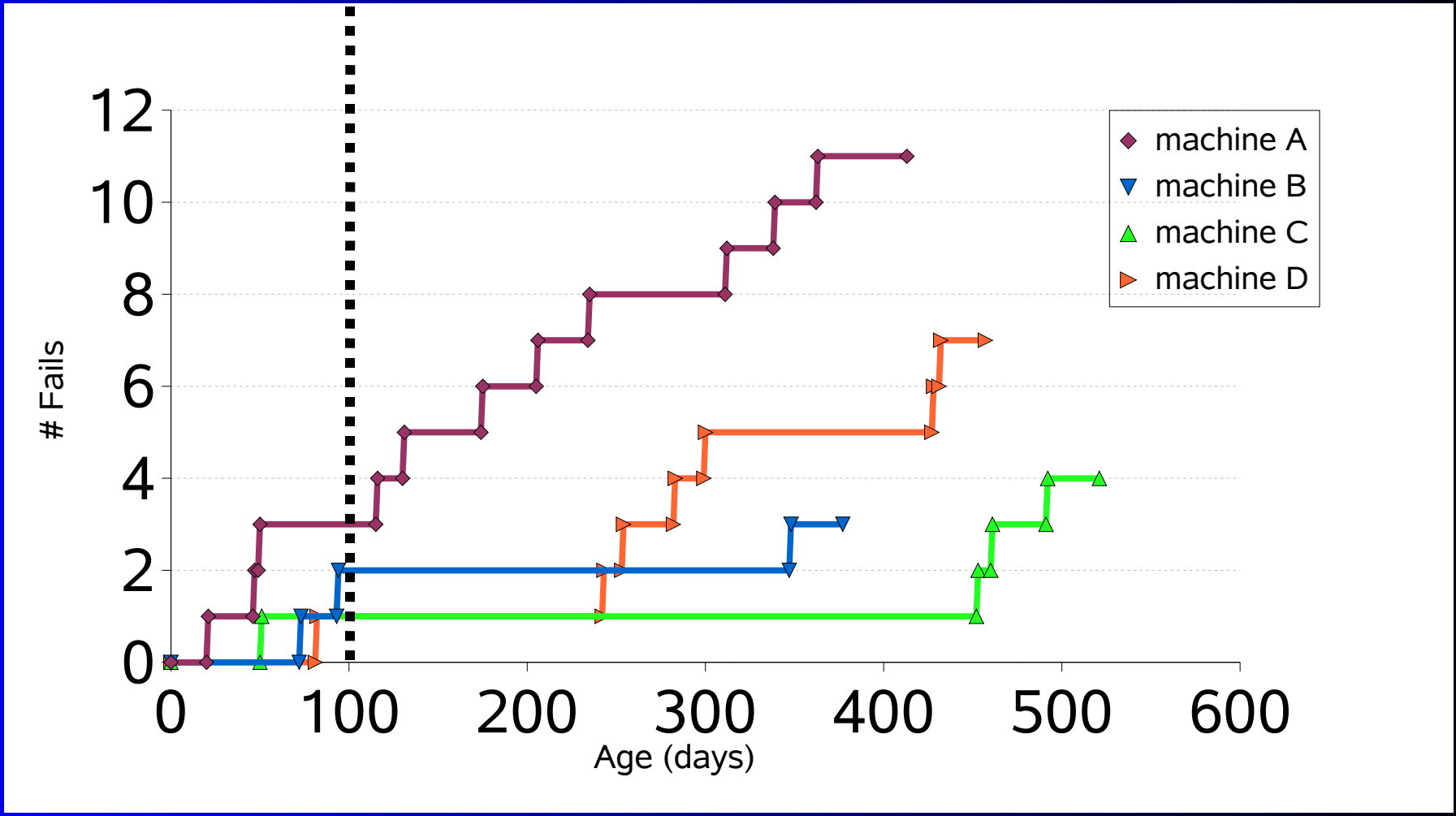


All 3 plots have 10 fails in 700 hours leading to an MTBF of 70 hours. Clearly these behaviors are different

Mean Cumulative Function (MCF)

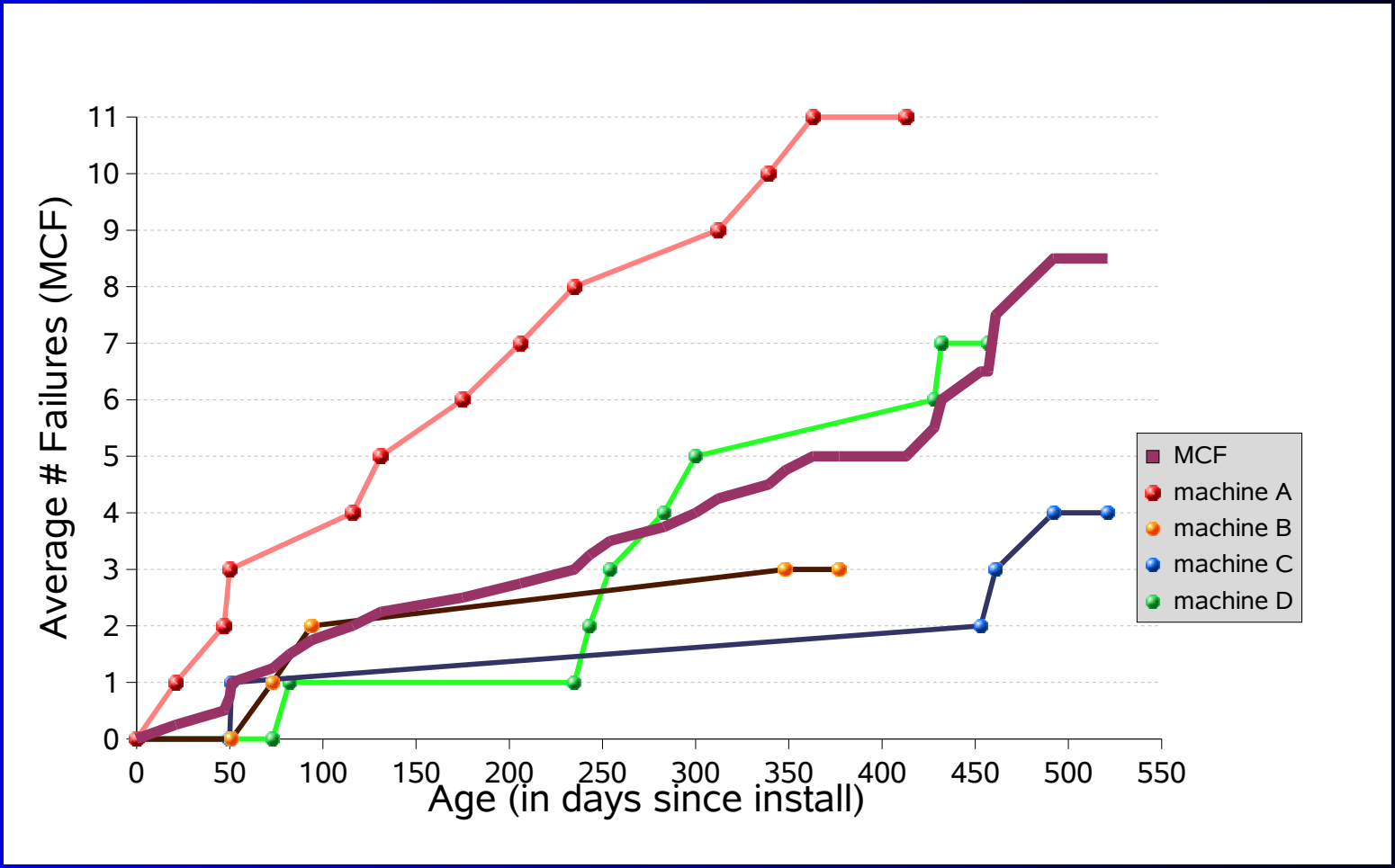
- MCF is the average of the cumulative plots across a group of systems at risk at any point in time.
- MCF is obtained from a vertical slice across the individual cumulative plots at a time point.
- MCF is the average number of failures of a group of systems at a particular age.

MCF from Cumulative Plots



Average at each vertical time slice is the MCF. We show slice at 100 hours on a collection of 4 cumulative plots.

MCF and Cumulative Plots

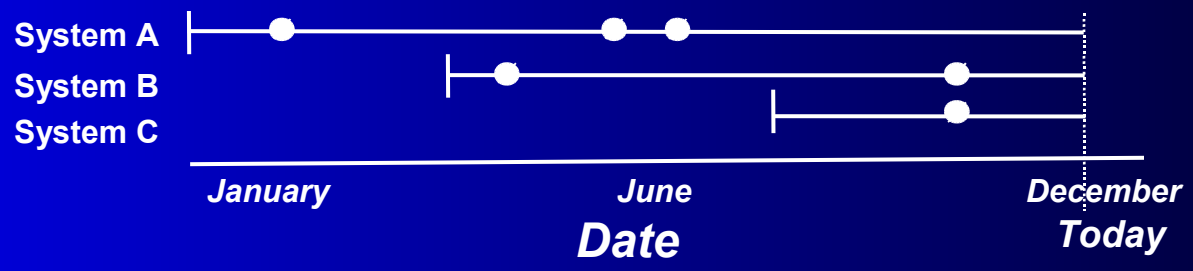


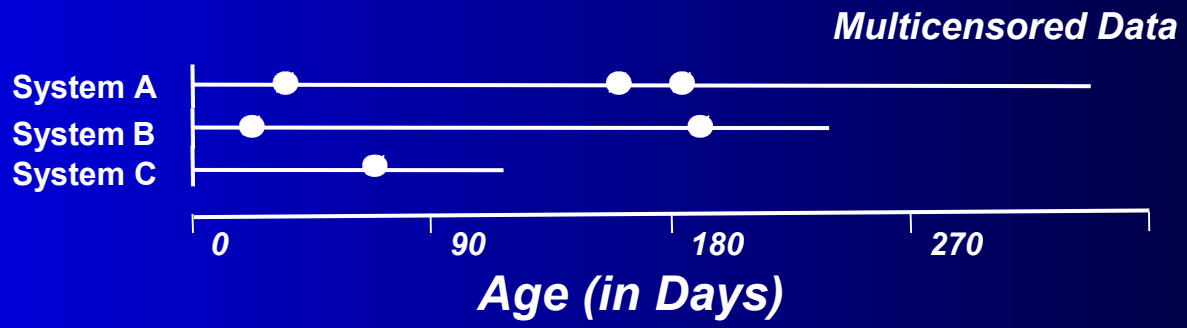
Note: Steps replaced with connecting lines

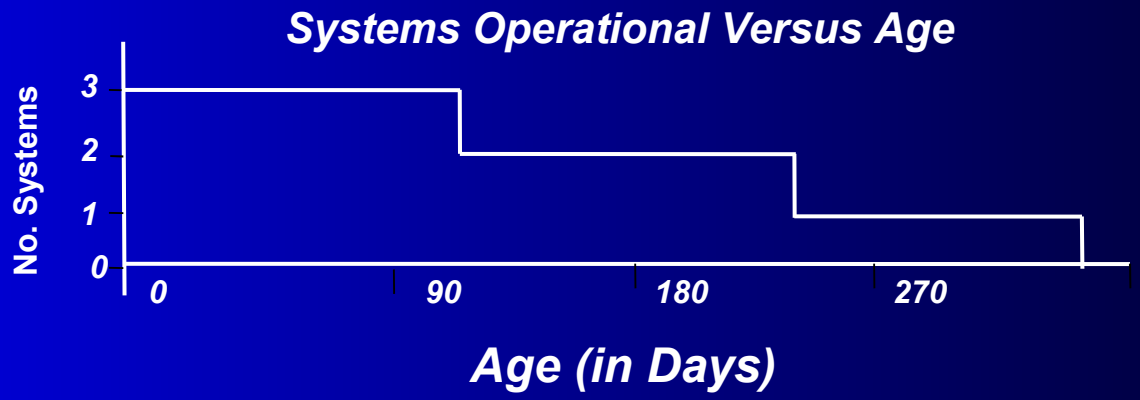
Multicensoring Issues

- Because systems are installed at different times throughout the year, system ages will differ resulting in ***multicensored data.***
- ***Right censored*** data has no failure information beyond a specific system age; e.g., if a machine is 100 days old it cannot contribute information regarding reliability at 150 days of operation.
- ***Left censored*** data has no information before a specific date; e.g., data collection begins on Jan 2004 and no failure history is available
- MCF accounts for the number of systems at risk at any age or date.

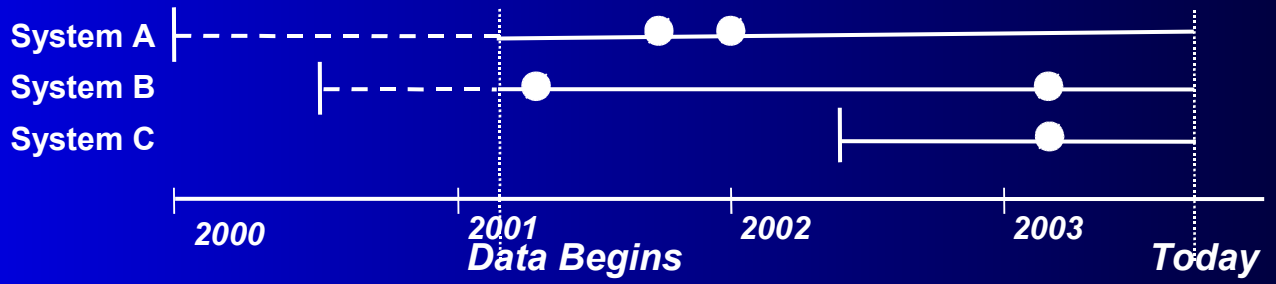
Right Censored Data

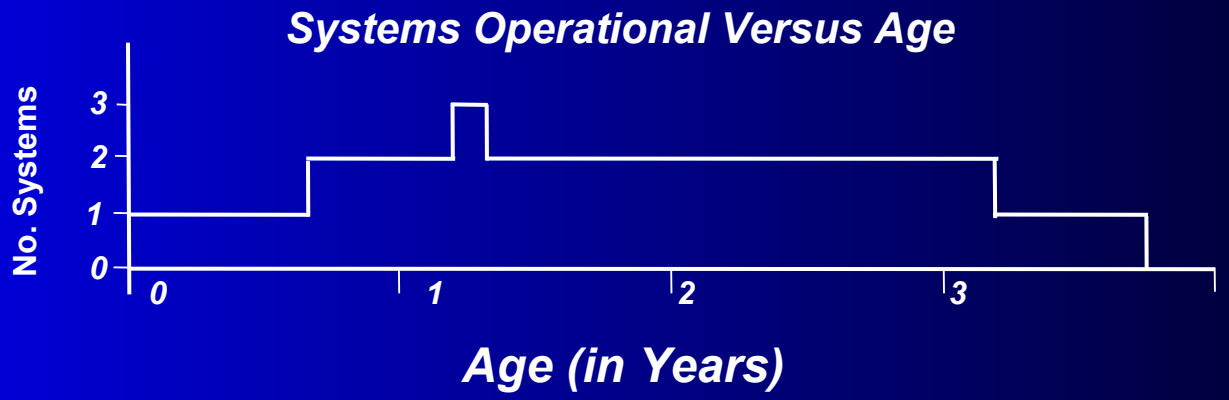




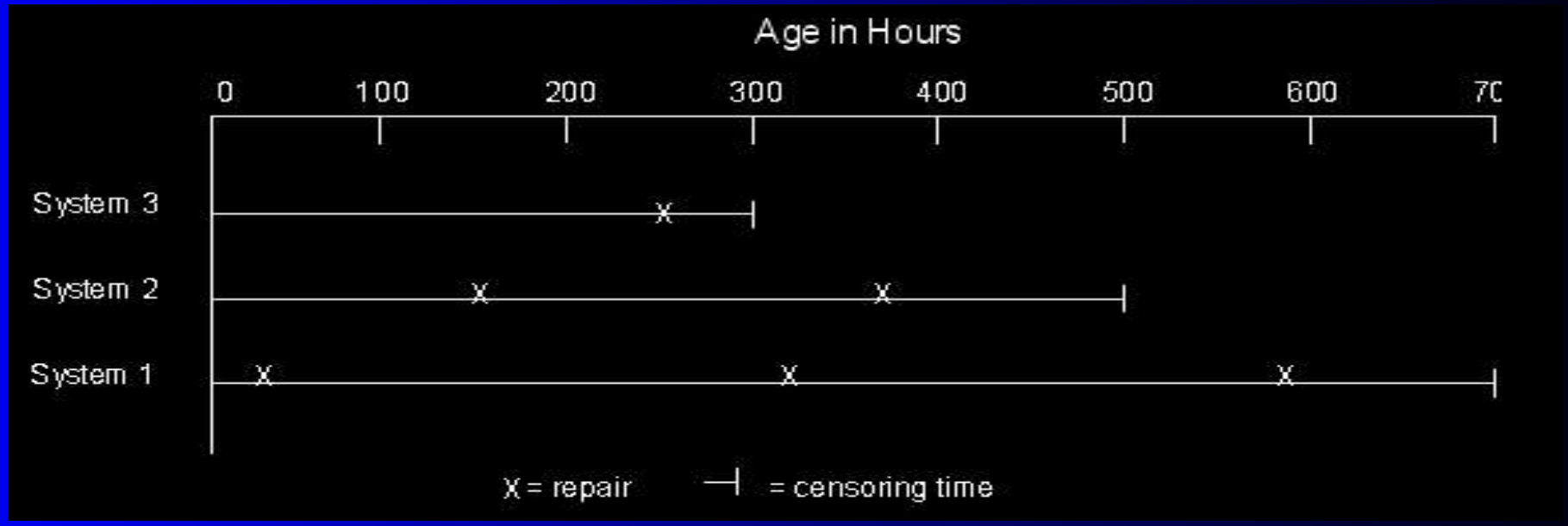


Left Censored Data





MCF Calculation (3 systems)



Time (Hrs)	33	135	247	300	318	368	500	582	700
Number of Systems at Risk	3	3	3	3	2	2	2	1	1
Fails/System	1/3	1/3	1/3		1/2	1/2		1/1	
MCF	1/3	2/3	3/3	3/3	3/3+1/2	3/3+2/2	3/3+2/2	3/3+2/2+1/1	3/3+2/2+1/1

Confidence bounds can be computed for MCF

MCF calculation : Data Template

- History on every machine, including systems without failures.
- Note install, begin, failure, and end event dates.

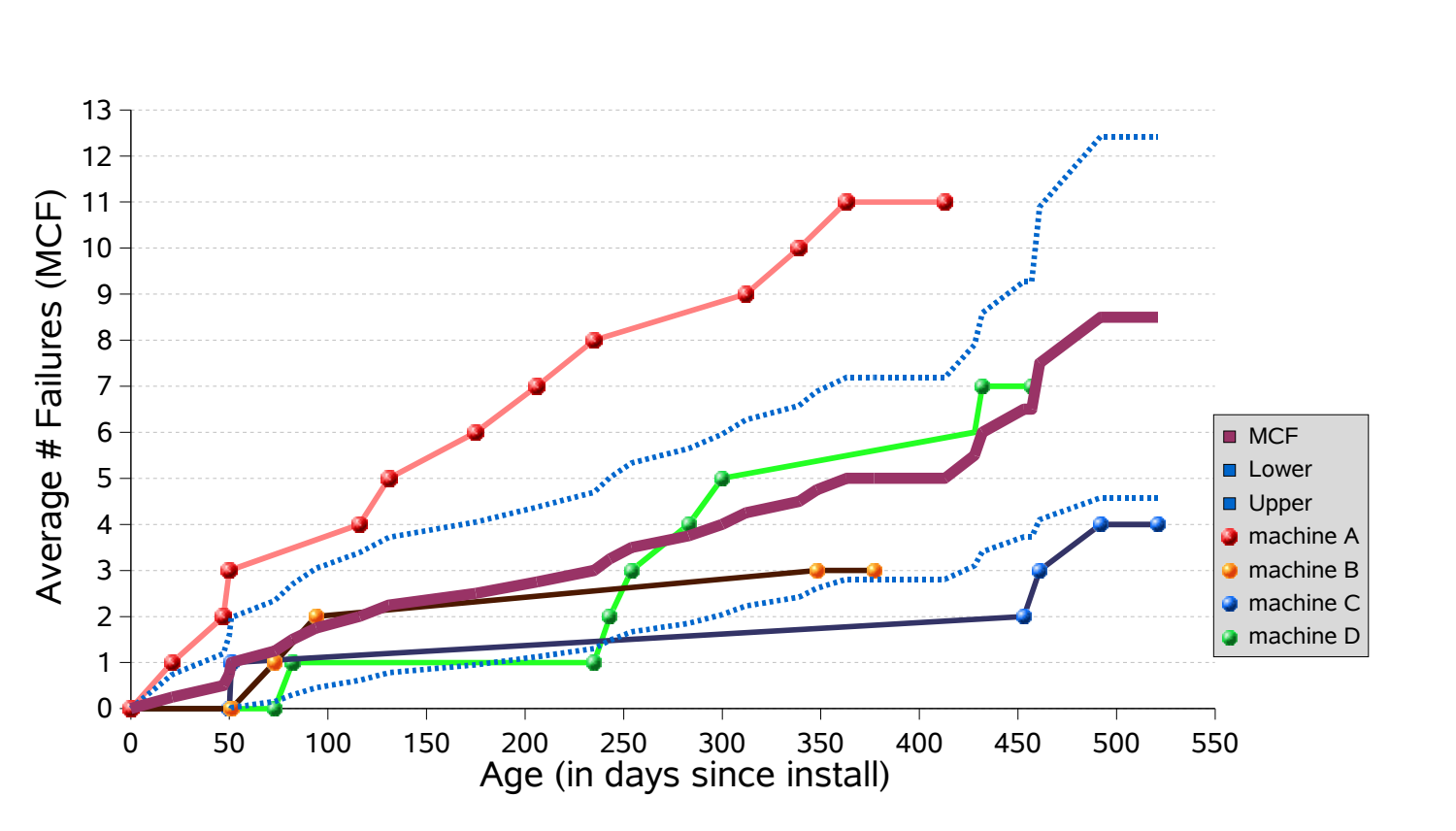
HostName	SN	Platform	Location	Install date	Event	Event Date	Event Age from install (days)	Case Id	Sys Impact	Type	Code	Cause
speedy1	555H5555	E3500	SALEM	8/11/2000	Install	8/11/2000	0					
speedy1	555H5555	E3500	SALEM	8/11/2000	begin	8/12/2000	1					
speedy1	555H5555	E3500	SALEM	8/11/2000	Failure	8/23/2001	378	55559945	I	HW	MEM	Memory Error
speedy1	555H5555	E3500	SALEM	8/11/2000	Failure	12/18/2001	495	55559984	I	SC	SC	System Config - improper sys. config.
speedy1	555H5555	E3500	SALEM	8/11/2000	Failure	3/1/2002	568	55559999	N	HW	PS	Power Supply/converter
speedy1	555H5555	E3500	SALEM	8/11/2000	end	8/12/2002	731					
speedy2	666F6666	E3500	SALEM	12/8/1997	Install	12/8/1997	0					
speedy2	666F6666	E3500	SALEM	12/8/1997	begin	8/12/2000	978					
speedy2	666F6666	E3500	SALEM	12/8/1997	end	8/12/2002	1708					

MCF calculation: Confidence Intervals

Age (i)	Event	Number at Risk (ni)	mi=1/ni	MCFi=MCFi-1 +mi	ci=mi/ni	V(MCFi)=vi=vi-1+ci	Lower Limit	Upper Limit
0	Begin	4		0		0	0	0
21	Fail	4	0.25	0.25	0.06	0.06	-0.24	0.74
47	Fail	4	0.25	0.5	0.06	0.13	-0.19	1.19
50	Fail	4	0.25	0.75	0.06	0.19	-0.1	1.6
51	Fail	4	0.25	1	0.06	0.25	0.02	1.98
73	Fail	4	0.25	1.25	0.06	0.31	0.15	2.35
82	Fail	4	0.25	1.5	0.06	0.38	0.3	2.7
94	Fail	4	0.25	1.75	0.06	0.44	0.45	3.05
116	Fail	4	0.25	2	0.06	0.5	0.61	3.39
131	Fail	4	0.25	2.25	0.06	0.56	0.78	3.72
175	Fail	4	0.25	2.5	0.06	0.63	0.95	4.05
206	Fail	4	0.25	2.75	0.06	0.69	1.12	4.38
235	Fail	4	0.25	3	0.06	0.75	1.3	4.7
243	Fail	4	0.25	3.25	0.06	0.81	1.48	5.02
254	Fail	4	0.25	3.5	0.06	0.88	1.67	5.33
283	Fail	4	0.25	3.75	0.06	0.94	1.85	5.65
300	Fail	4	0.25	4	0.06	1	2.04	5.96
312	Fail	4	0.25	4.25	0.06	1.06	2.23	6.27
339	Fail	4	0.25	4.5	0.06	1.13	2.42	6.58
348	Fail	4	0.25	4.75	0.06	1.19	2.61	6.89
363	Fail	4	0.25	5	0.06	1.25	2.81	7.19
377	End	3		5		1.25	2.81	7.19
413	End	2		5		1.25	2.81	7.19
428	Fail	2	0.5	5.5	0.25	1.5	3.1	7.9

Confidence limits based on Nelson's book.

MCF, CI, Cumulative Plots



- In one year of operation this population has 5 fails/machine with an upper bound of 7.
- Machine A clearly has higher than average failures at all ages.

Detecting anomalous machines

Naïve Confidence Intervals (from Nelson)

- If machine cumulative function is well above the upper bound it is deemed anomalous.
- Method works well on small sample sizes (eyeball approach)
- Graphically focuses attention on machines with high failures (over all or in small time windows).
- Prediction intervals on MCF will be more accurate.

Glosup's Method

- One machine is removed from the population and the MCF is computed.
- MCF with N machines is compared with MCF with (N-1) machines for all combinations.
- Anomalous machine is based on the difference between these MCF combinations.
- See reference.

Heavlin's Method

- Based on Cochran-Mantel-Hanzel statistic for 2X2 contingency tables.
- Powerful but computationally intensive
- To be published.

Recurrence Rate

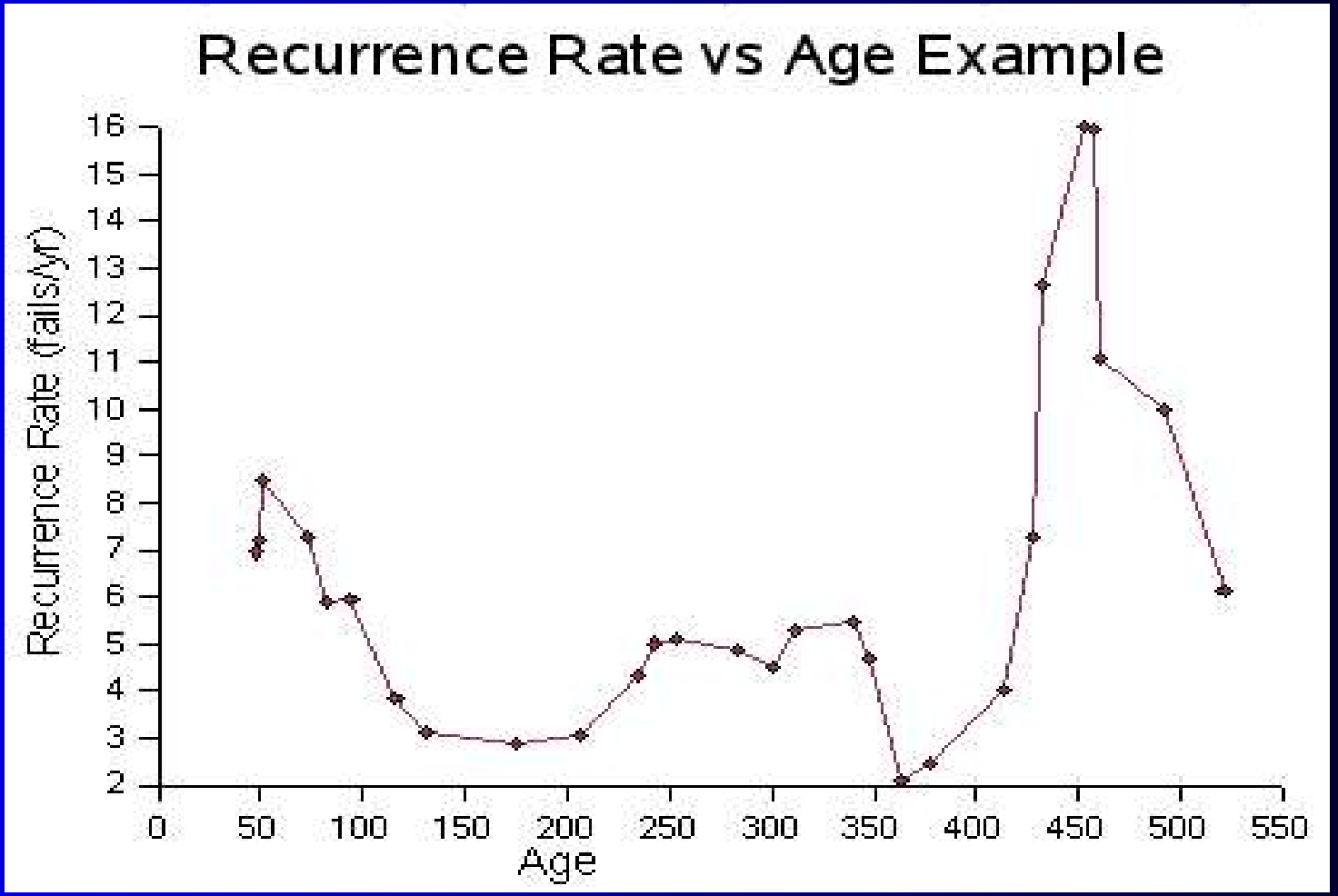
- A key characteristic of the MCF is the recurrence rate determined from the slope of the MCF at any point in time.
- The local slope represents the rate at which failures are occurring.
- Local slope is estimated by fitting a line to a group of points in a “window”.
- Degree of smoothing is the number of points used in estimating the tangent to the MCF.
- Slope(X,Y) functions in spreadsheets can be used to obtain recurrence rates easily.

Recurrence Rate : Step by Step

Age (i)	Event	Number at Risk (ni)	mi=1/ni	MCFi=MCFi-1 +mi	Recurrence Rate =Slope(MCF vs Age)
0	Begin	1			
0	Begin	2		0	
0	Begin	3		0	4.35
0	Begin	4		0	3.92
21	Fail	4	0.25	0.25	4.76
47	Fail	4	0.25	0.5	5.86
50	Fail	4	0.25	0.75	7.21
51	Fail	4	0.25	1	8.48
73	Fail	4	0.25	1.25	7.29
82	Fail	4	0.25	1.5	5.9
94	Fail	4	0.25	1.75	5.94
116	Fail	4	0.25	2	3.85
131	Fail	4	0.25	2.25	3.13
175	Fail	4	0.25	2.5	2.87
206	Fail	4	0.25	2.75	3.06
235	Fail	4	0.25	3	4.34
243	Fail	4	0.25	3.25	5.02
254	Fail	4	0.25	3.5	5.11
283	Fail	4	0.25	3.75	4.85
300	Fail	4	0.25	4	4.49
312	Fail	4	0.25	4.25	5.29
339	Fail	4	0.25	4.5	5.5
348	Fail	4	0.25	4.75	4.66
363	Fail	4	0.25	5	2.11
377	End	3		5	2.45
413	End	2		5	4
428	Fail	2	0.5	5.5	7.28

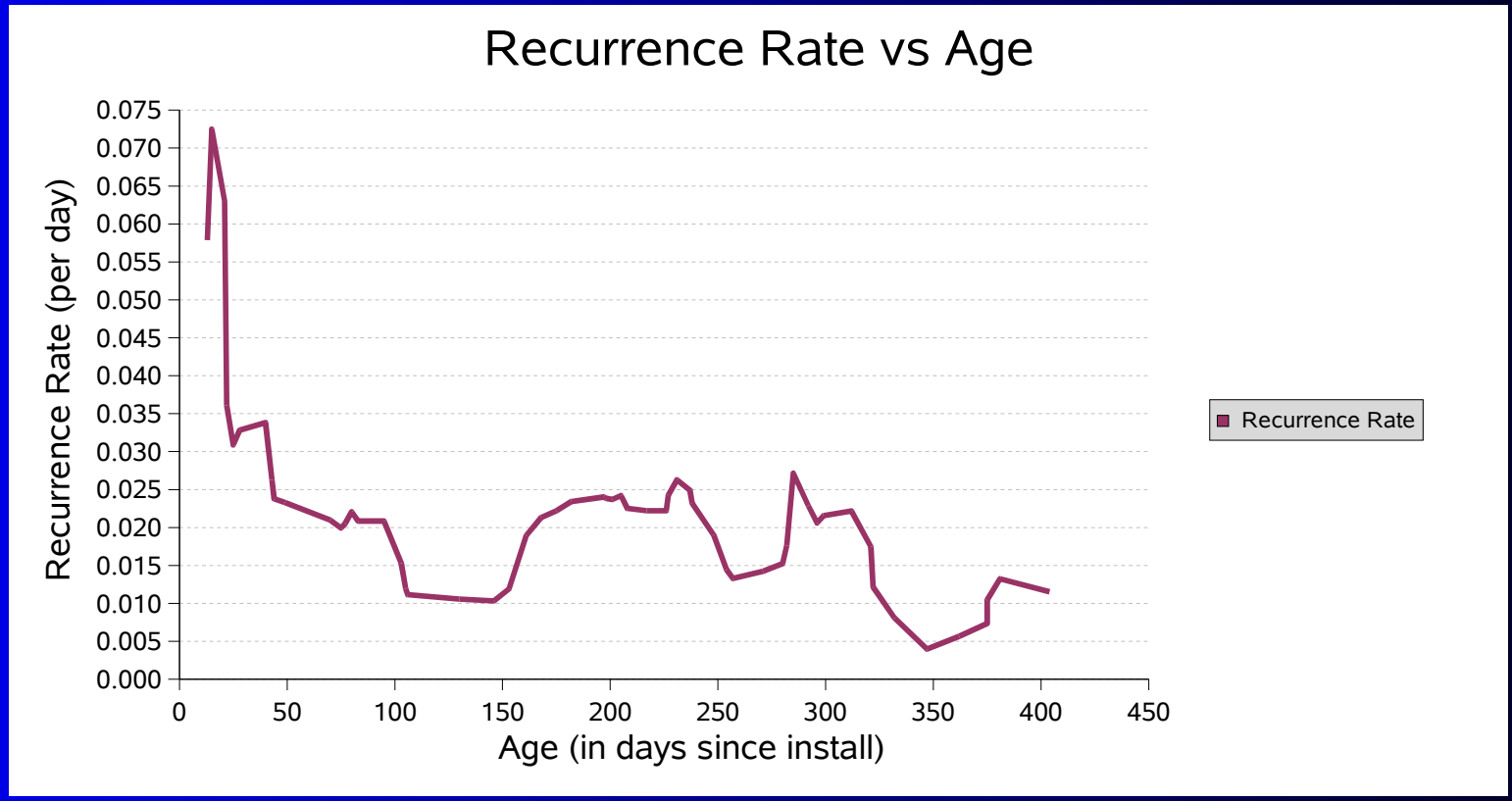
5 point slope

Recurrence Rate Vs Age



Recurrence rate peak at age 450 caused by single system.

Recurrence Rate Vs Age Example

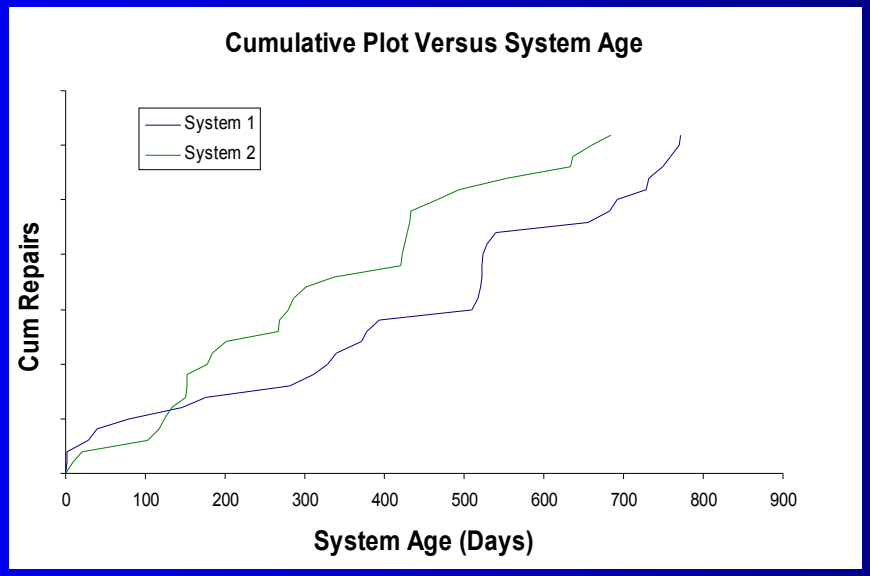


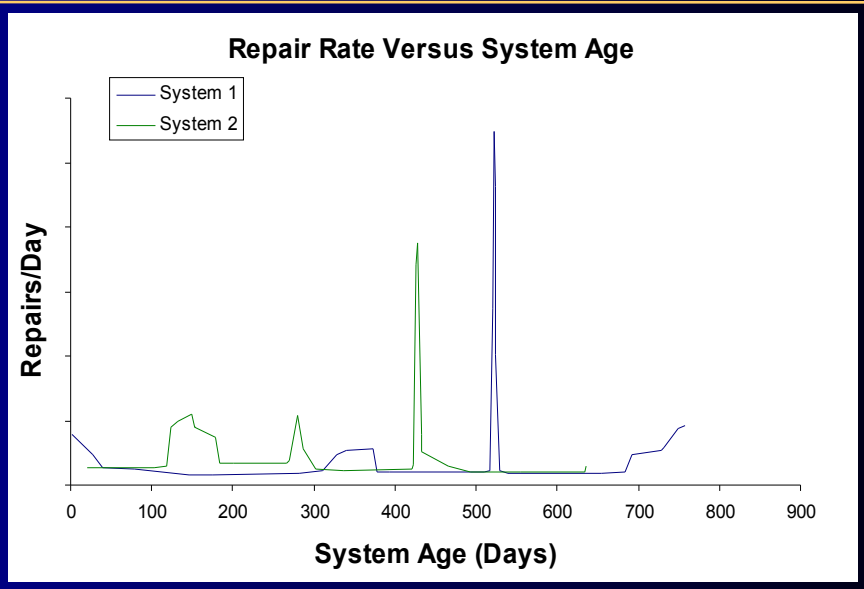
- Clearly rate of failures is decreasing with age.
- “Peaks” or “spikes” are related to multiple fails in short time periods indicating possible misdiagnosis or dead on arrival (DOA) spares.

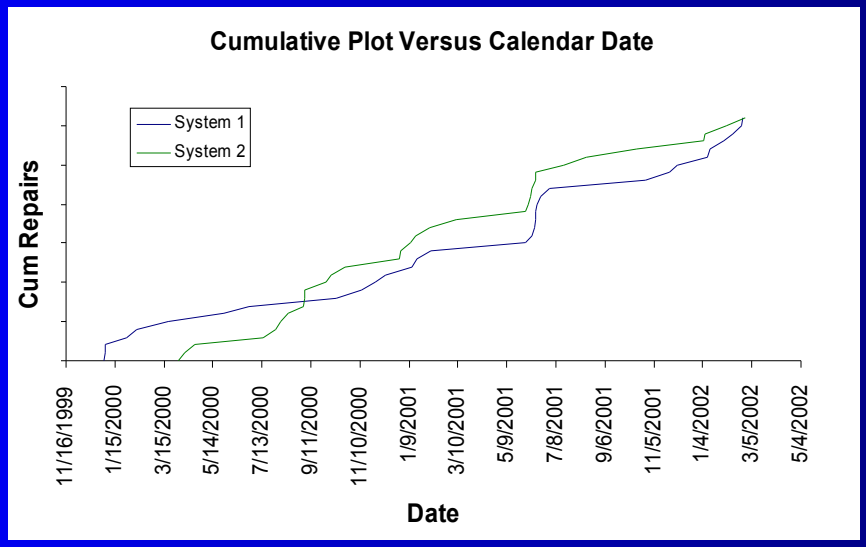
Calendar Time Analysis

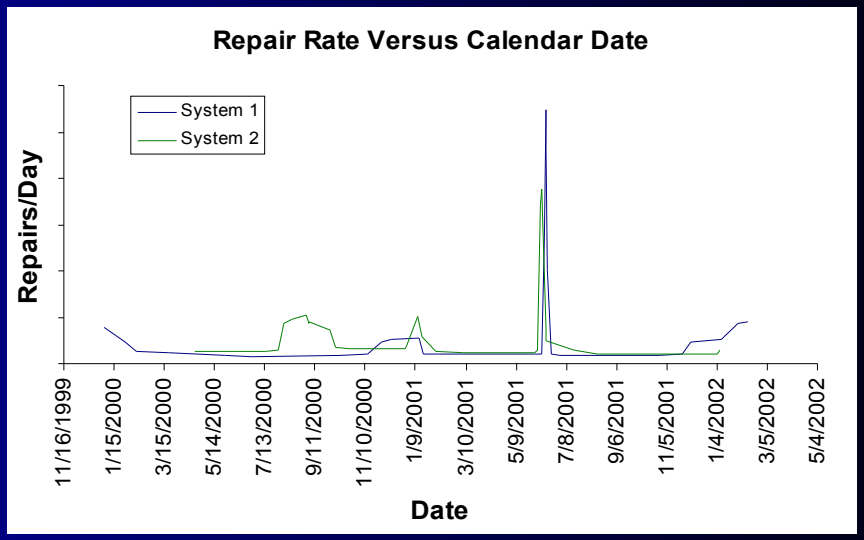
- Applications of patches, upgrades to faster processors, changing of operational policy of systems administrators, etc., affect a population of machines which are at several ages on a single date or calendar window.
- These effects are captured by MCF and recurrence rates versus calendar time instead of age.
- Slope(X,Y) in spreadsheets handles both age (numbers) or dates. Dates are automatically converted to days elapsed.

Calendar Time Analysis



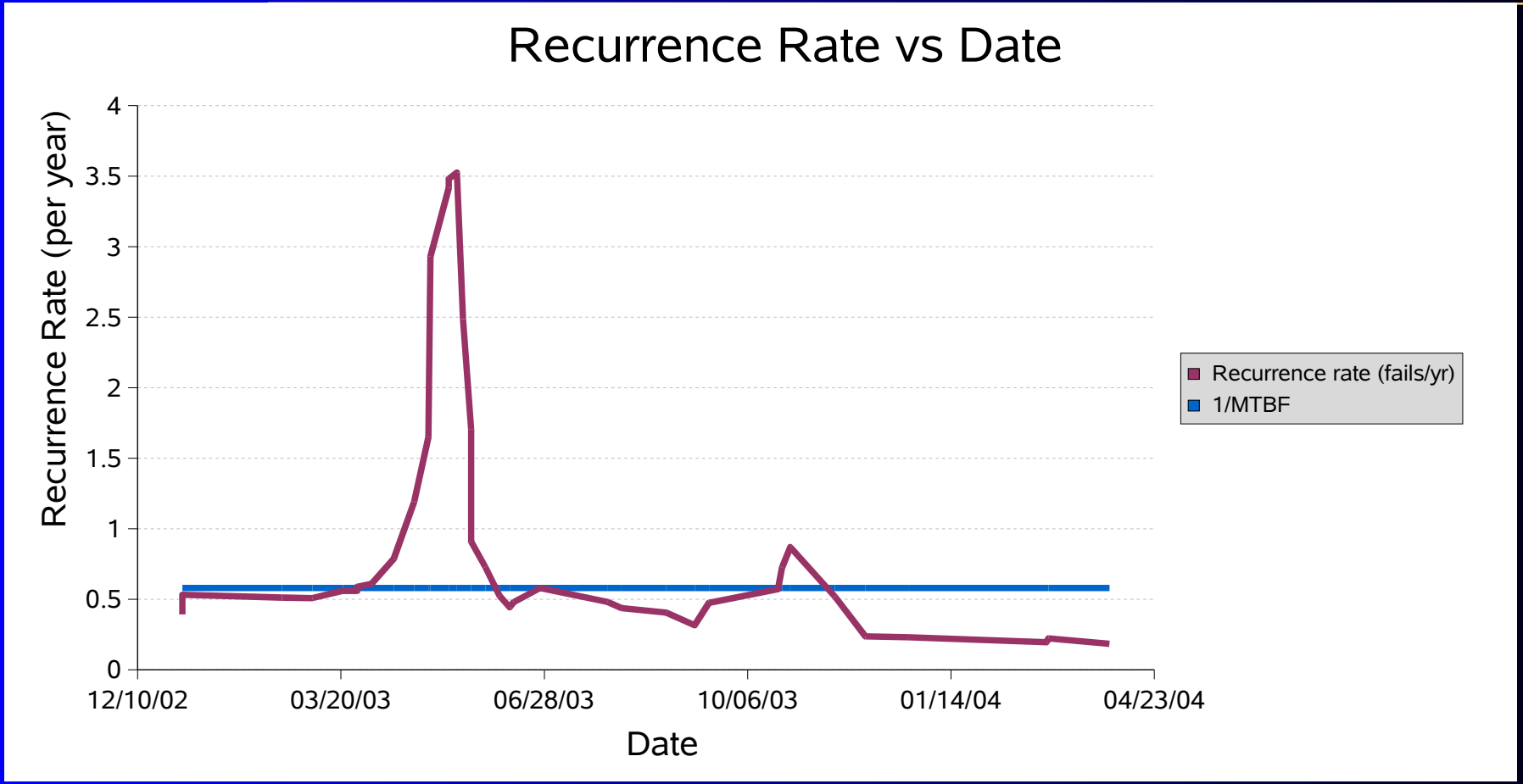






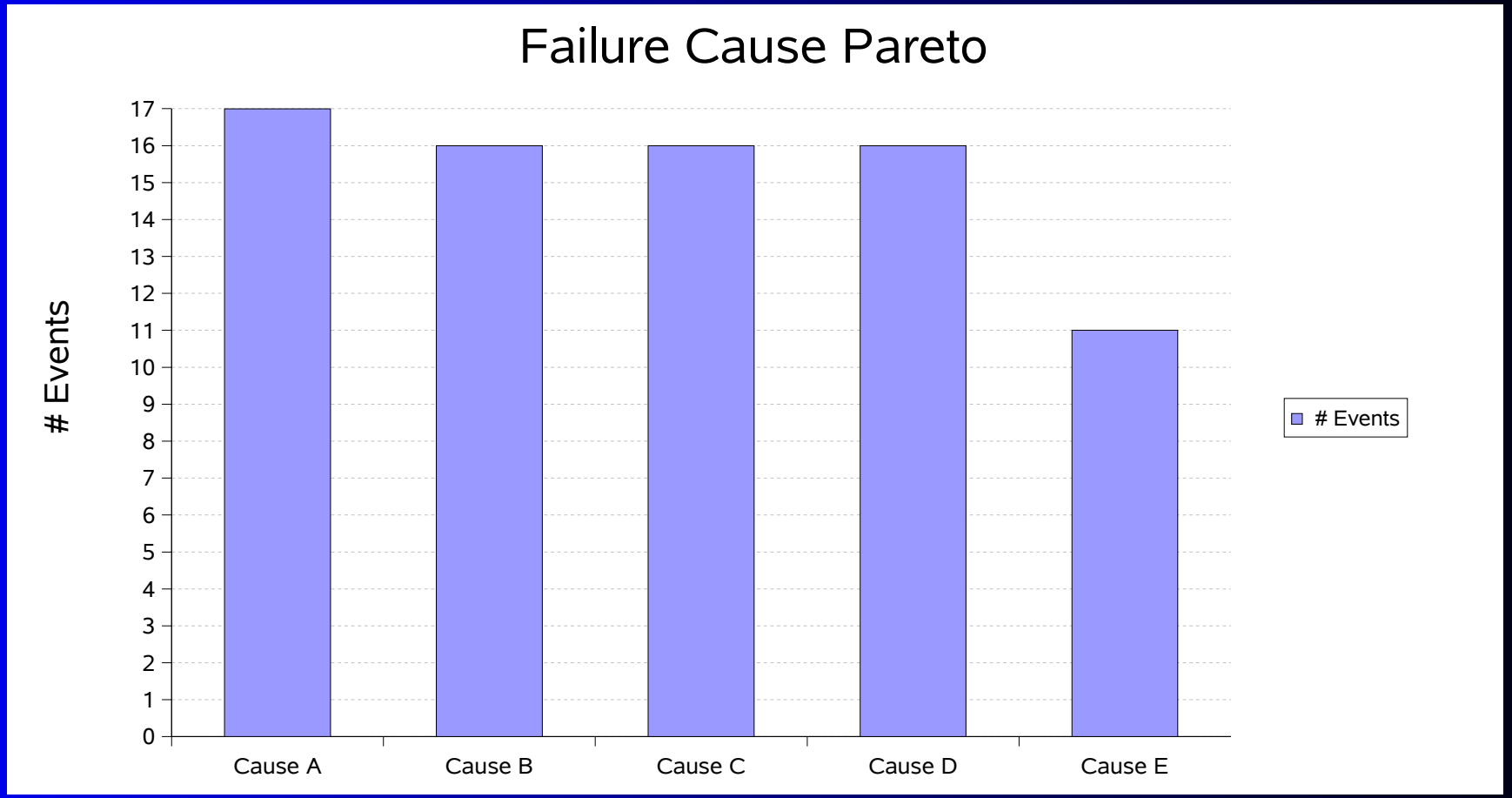
Simulated data: Two systems. One installed 1/1/2001and second 4/1/2001. Software upgrades on both systems on 6/1/2001.

Recurrence Rate Vs Date Example



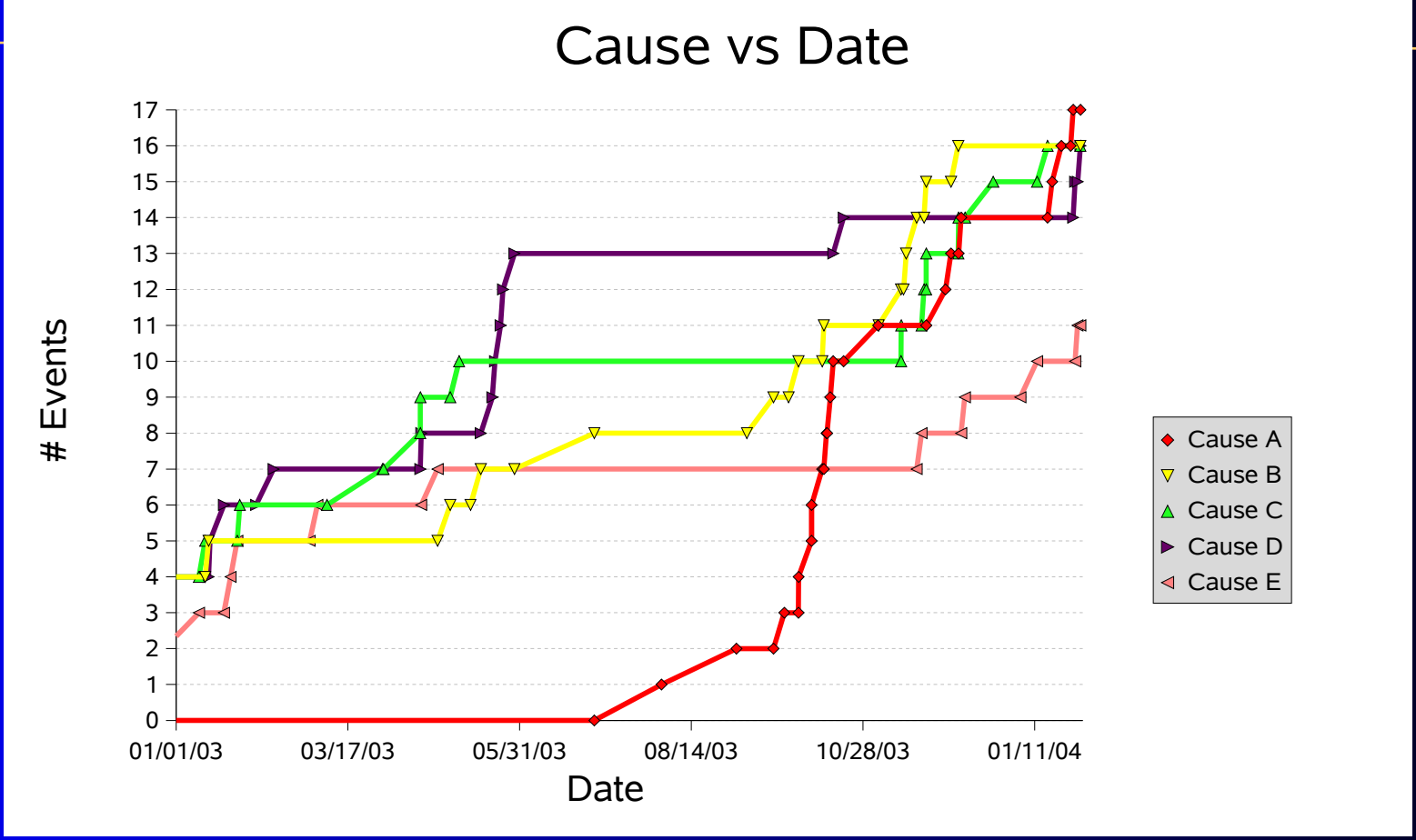
- Sharp increase in rate of fails in Apr-Jun followed by stable rate
- Rate decreases by half from January.
- Spike was related to multiple fails on a single machine in a short time period.

Failure Cause Pareto



- Pareto charts are static.
- Plot does not show which causes have been remediated and which ones are current threats.

Failure Cause Plots

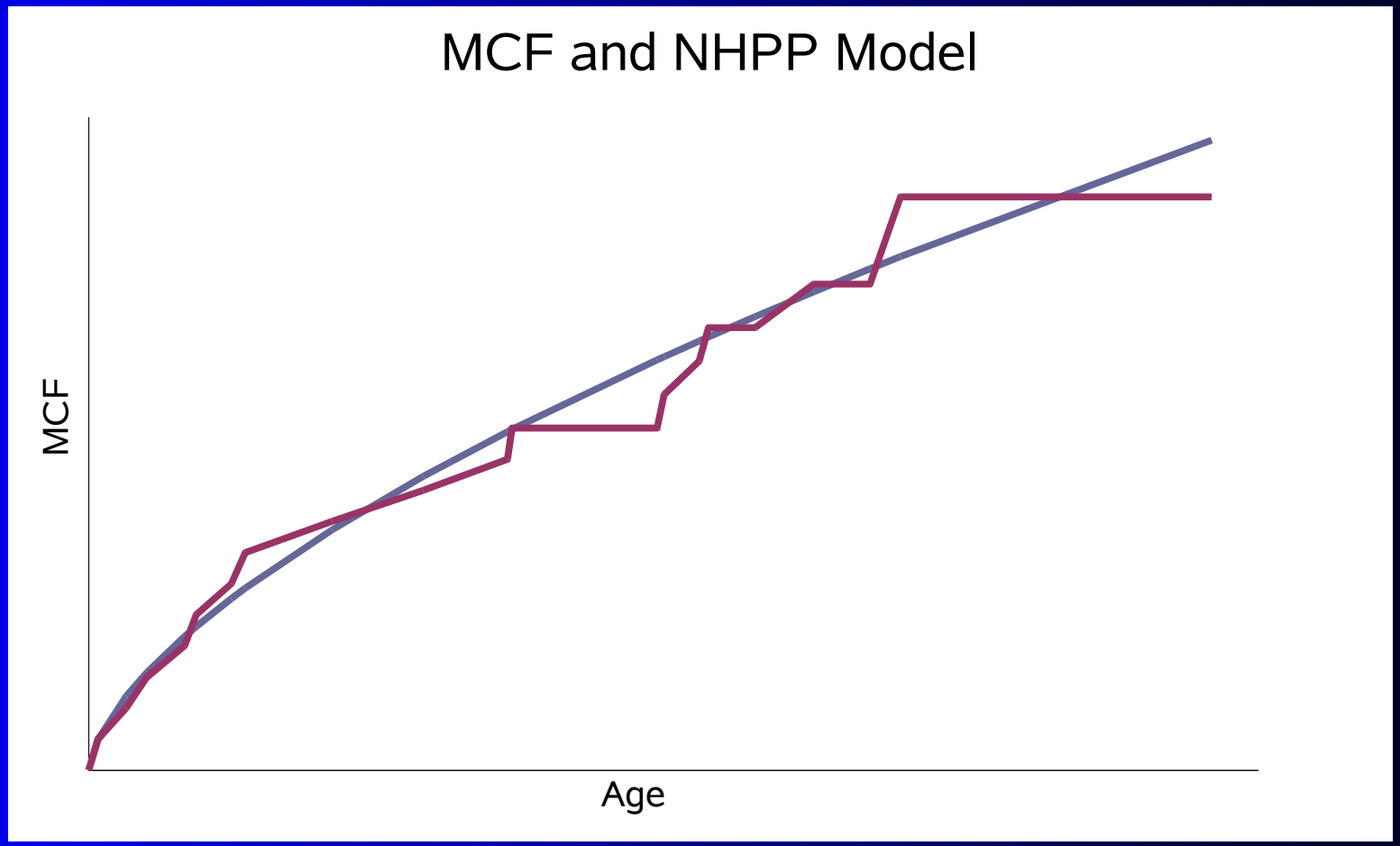


- Failure cause can be plotted against age or date.
- Cause D was remediated in '03 while Cause A is a current threat.
- Conveys time evolution of customer problems.

Additional uses of MCFs

- MCF Vs system age can be used to compare various (sub) populations despite multicensoring.
  - Machines at different customer sites.
  - Machines belonging to the same customer but located at different datacenters.
  - Machines of different vintages e.g., manufactured in 2003 Vs 2004.
  - Machines performing different functions e.g., production Vs development.
- Case studies will illustrate these uses.

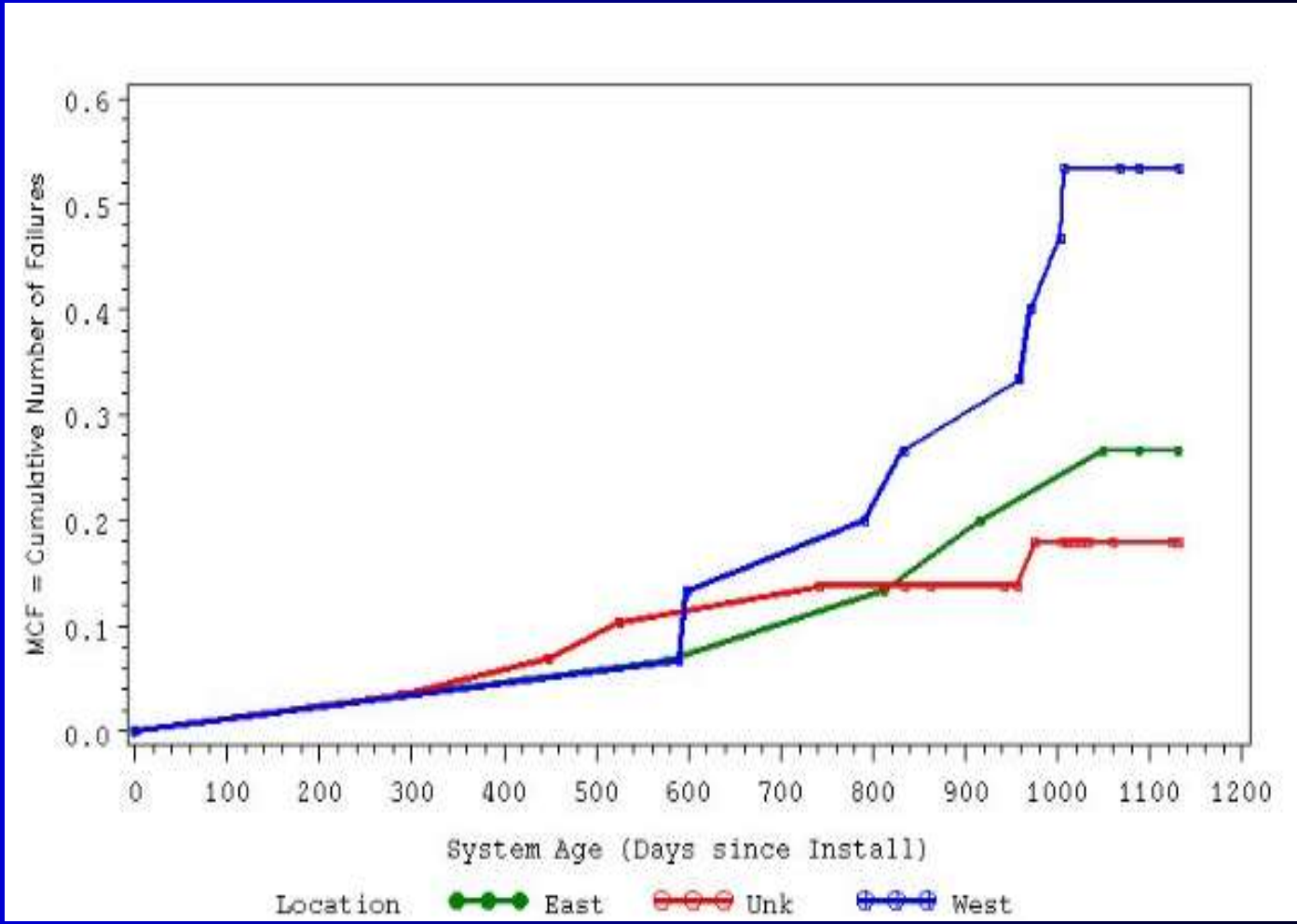
NHPP Fitting



- Example of fitting a power law model to an MCF
- The parametric fit to a non parametric MCF can be used for prediction and extrapolation.

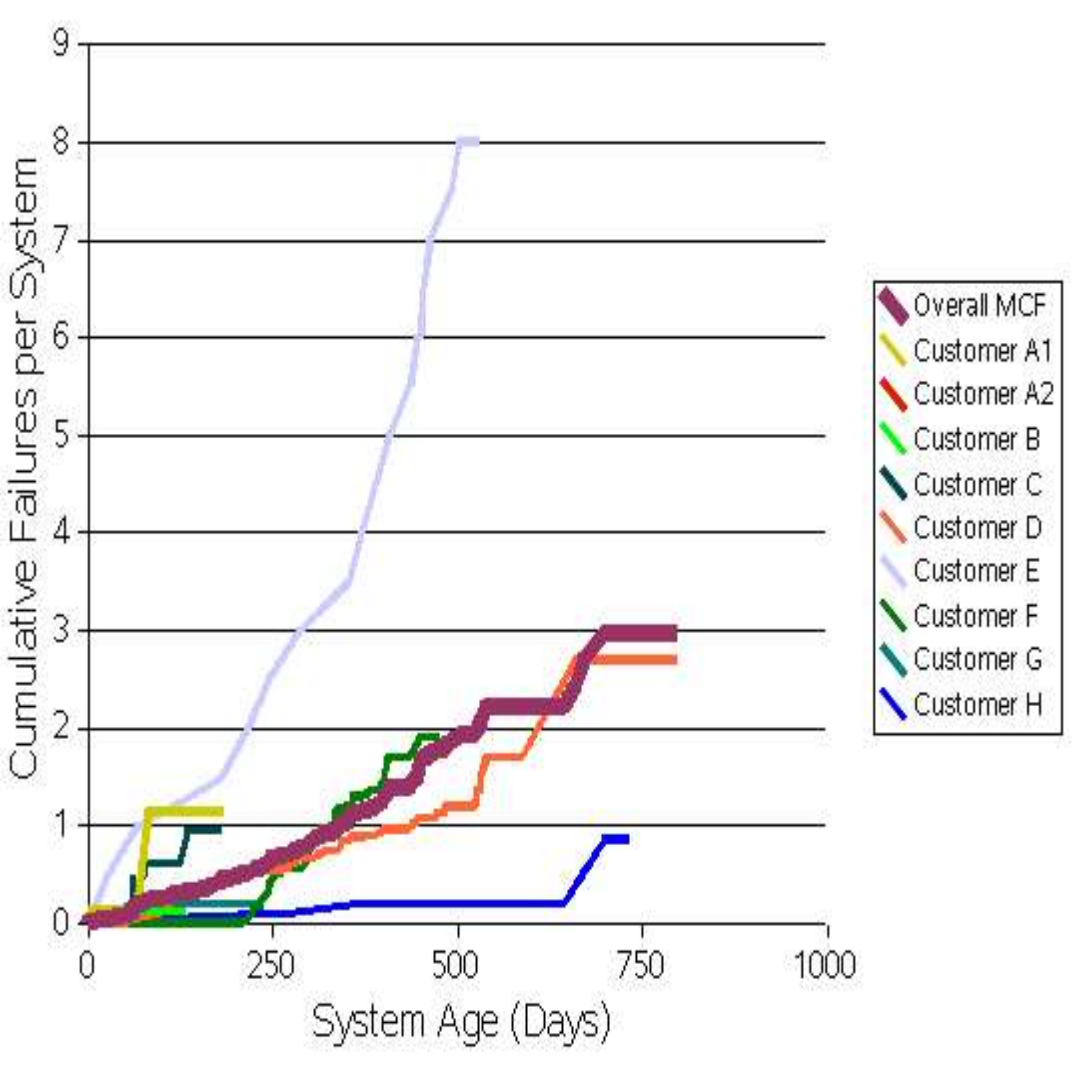
Case Study 1

Customer with East Coast Vs West Coast data centers. Air conditioning issues caused West coast problems.



Case Study 2

Comparison across customers, common platform.



Transforming Sun

- MTBF represented the comfort zone.
- Struggled to push TDR until successes achieved at various customers in resolving reliability issues.
  - “You guys have hit a home run.”
  - “Some of the best work I've seen from Sun.”
- Training of Sun engineers and field service personnel in TDR usage is ongoing.
- Sun has developed software tools for internal use to facilitate analysis and generate reports.

Summary

- The analysis of repairable systems does not have to be complicated.
- There are better ways to measure reliability than just MTBF.
- Measuring and monitoring repairable system reliability can be greatly facilitated by the use of simple but very powerful graphical techniques.

Benefits

- These procedures have been applied very effectively at Sun Microsystems for monitoring the reliability of Sun equipment at many customers.
- These methods have allowed Sun engineers to quickly identify trends, anomalous systems, unusual behavior, the effects of hardware and software changes, maintenance practices, and installation actions.
- Customers presented with TDR analysis reports have responded very favorably to these revealing charts.

Where to Get More Information

Glosup, J., “Detecting Multiple Populations within a Collection of Repairable Systems”, Joint Statistical Meetings, Toronto, Canada, 2004

Lawson, J.S., Wesselmann, C.W., Scott, D.T.,"Simple Plots Improve Software Reliability Prediction Models", *Quality Engineering*, Vol 15. No. 3. pp411-417, 2003.

Nelson W., “Graphical Analysis of System Repair Data”, *Journal of Quality Technology*, **17**, 140-146.

Nelson, W., *Recurrence Events Data Analysis for Product Repairs, Disease Recurrences and Other Applications*, ASA-SIAM series on Statistics and Applied Probability, 2003.

Tobias, P.A., Trindade, D.C., *Applied Reliability*, Chapman & Hall/CRC, 1995.

Trindade, D.C., “An APL Program to Numerically Differentiate Data“, IBM TR Report 19.0361, January 12, 1975

Trindade, D.C., Nathan, S., “Simple Plots for Monitoring the Field Reliability of Repairable Systems”, *Annual Reliability and Maintainability Symposium*, Alexandria, Virginia, 2005

Usher, J.S., "Case Study: Reliability Models and Misconceptions“, Quality Engineering, Vol. 6, No. 2, pp 261-271.

Presenter's Biographical Sketch

Dr. David Trindade is a Distinguished Engineer at Sun Microsystems. Formerly he was a Senior Fellow at AMD. His fields of expertise include reliability, statistical analysis, and modeling of components, systems, and software, applied statistics, especially design of experiments (DOE), and statistical process control (SPC). He is co-author (with Dr. Paul Tobias) of the book *Applied Reliability*, 2nd ed., published in 1995. He has authored many papers and presented at many international conferences. He has a BS in physics, an MS in material sciences and semiconductor physics, an MS in statistics , and a Ph.D. in mechanical engineering and statistics. He has been an adjunct lecturer at the University of Vermont and Santa Clara University.

e-mail: david.trindade@sun.com

Dr. Swami Nathan is a Senior Staff Engineer in Sun Microsystems. His field of interest is field data analysis, statistical analysis and reliability/availability modeling of complex systems. He received his B.Tech from Indian Institute of Technology, and M.S. and Ph.D. in reliability engineering from the University of Maryland, College Park. He has authored over a dozen papers and presented at international conferences and holds two patents.

email: swami.nathan@sun.com