

IAF0530/IAF9530

Süsteemide usaldusväärsus ja veakindlus Dependability and fault tolerance

Gert Jervan
Department of Computer Engineering
Tallinn University of Technology

General Information

- Contents:
Dependability and fault tolerance
www.pld.ttu.ee/IAF0530
- Lecturer & Examiner:
Gert Jervan
IT-229 620 2261
gert.jervan@pld.ttu.ee
www.pld.ttu.ee/~gerje

2

Course Plan

- 16 occasions, á 1,5 hours
Tuesdays 12:00-13:30
- 8 Lectures (No meetings on Feb 28, March 13, May 8)
- Case Studies
 - Introductory presentation (5 min)
 - 20 min/30 min presentation of the final report
 - Written report (6 pages, using predefined template; 10 pages for PhD students)
- Exam (sort of)

3

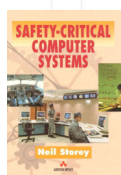
Reading

- **Various papers (on the course homepage)**
www.pld.ttu.ee/IAF0530
- Textbooks
- Incident/accident reports
- Web pages

4

Textbooks

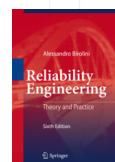
- Safety-critical Computer Systems
 - Neil Storey
 - Addison Wesley, 1996.
 - An introductory text which provides overview of safety related aspects and methods in computer systems development.



5

Textbooks

- Reliability Engineering: Theory and Practice.
 - Alessandro Birolini
 - Springer
 - 2010 (6th ed.), 2007 (5th ed.)
 - This book shows how to build in, evaluate, and demonstrate reliability & availability of components, equipment, systems. It presents the state-of-the-art of reliability engineering, both in theory and practice



6

Textbooks

- Fault-Tolerant Systems
 - Israel Koren and C. Mani Krishna
 - Morgan-Kaufman Publishers, 2007



This book covers comprehensively the design of fault-tolerant hardware and software, use of fault-tolerance techniques to improve manufacturing yields and design and analysis of networks. Additionally it includes material on methods to protect against threats to encryption subsystems used for security purposes.

7

Case Studies

- Topic categories:
 - Accident analysis
 - System safety analysis
 - Literature survey
 - Something else (implementation, tool study, etc.)
- Requires prior ack.

Literature and sample (!) topics on the webpage

8

Case Studies

- Some examples:
 - Clock synchronization
 - Atomic and reliable broadcast
 - Algorithmic based fault-tolerance
 - System level diagnosis - distributed algorithms
 - Fault-tolerant transaction processing systems
 - Measures of software reliability
 - Validation and verification techniques
 - CAN (Controller Area Network) protocol
 - Fault-Tolerance in E-Commerce Web Servers
 - Fault tolerance in wired and wireless systems
 - Nano tubes
 - ...

9

Case Studies

- Topic selection:
 - February 28 (via e-mail, no lecture at that day)
- Draft of the report (incl. introductory presentation of the topic, 1 page):
 - March 20
- Presentations:
 - April 24 – May 15
- Final Report:
 - May 31st
 - The best reports will be published in A&A (selected topics only, *preferably* in Estonian)

10

Course Outline (Preliminary)

- Feb 7: Introduction
- Feb 14, Feb 21: Lectures II and III
- Feb 28: No lecture – topic selection (via e-mail)
- March 6: Lecture IV
- March 13: No lecture
- March 20: Introductory presentation of the topic (5 min each)
- March 27 – April 17: Lectures V - VIII
- April 24 – May 15: Case study presentations

11

Course overview

12

Course Overview

- Reliability: increasing concern
 - Historical
 - High reliability in computers was needed in critical applications: space missions, telephone switching, process control etc.
 - Contemporary
 - Extraordinary dependence on computers: on-line banking, commerce, cars, planes, communications etc.
 - Hardware is increasingly more fault-prone (complexity, technology, environment)
 - Software is increasingly more complex
 - Things simply do not work without special reliability measures

13

Hardware - Background

- Chip designers, device engineers and the high-reliability community recognize that reliability concerns ultimately limit the scalability of any generation of microelectronics technology
- Statistical methods and reliability physics provide the foundation for better understanding the next generation of scaled microelectronics
 - Microelectronics device physics
 - Reliability analysis and modeling
 - Experimentation
 - Accelerated testing
 - Failure analysis
- The design, fabrication and implementation of highly aggressive advanced microelectronics requires expert controls, modern reliability approaches and novel qualification strategies

14

Scaling Trends & Reliability Considerations

- A lot of technology concerns:
 - Reduced gate oxide thicknesses
 - Increased thermal/power densities
 - Reduced interconnect dimensions
 - Higher device operating temperatures
 - Increased sensitivity to defects and statistical process variations
 - Introduction of new materials with each new generation, replacing proven materials
 - e.g. Cu and low K inter-level dielectrics for Al and SiO₂

15

Scaling Trends & Reliability Considerations

- Dramatic increase in processing steps with each new generation
 - approx. 50 more steps per generation and a new metal level every 2 generations
- Rush to market - Less time to characterize new materials than in the past
 - e.g. reliability issues with new materials not fully understood and potential new failure modes
- Manufacturers' trends to provide 'just enough' lifetime, reliability, and environmental specs for commercial & industrial applications
 - e.g. 3-5 yr product lifetimes, trading off 'excess' reliability margins for performance

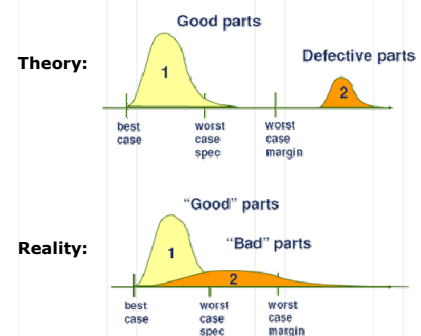
16

Scaling Trends & Reliability Considerations

- Significant rise in the amount of proprietary technology and data developed by manufacturers, reluctance to share information with hi-relevance customers
 - e.g. process recipes, process controls, process flows, design margins, MTTF
- Next generation microelectronics focus on the performance needs of the commercial customer, with little or no emphasis on the extreme needs
 - e.g. extended life, extreme environments, high reliability
- Increasingly difficult testability challenges due to device complexity

17

Correct or Defective?



18

Product Technical Trends

	1990	2000	2010
Operating temperature, °C	-55 to 125	-40 to +85	0 to 70
Supply voltage	5v	1.5v	0.6v
Max. power (high perf.)	5	100	170
No. of package types	<10	<80	??
Design support life	>10 yrs.	1-5 yrs.	<1yr.
Production life	>10 yrs.	3-5 yrs.	<3yrs.
<u>Service life</u>	<u>>20 yrs.</u>	<u>5-10 yrs.</u>	<u><5yrs.</u>

*MRQW-2002, Bernstein

19

Software complexity is a challenge

Aviation:

- Boeing 747 → 0.4 M LOC
- Boeing 777 → 4 M LOC
- Technology Review 2002

Automotive:

- ✓ 2010 Premium → 100 M LOC
- ✓ 1995 – 2000 → 52%/Year
- ✓ 2001 – 2010 → 35%/Year
Tony Scott, GM CIO
- ✓ 2011 – BMW is the first manufacturer to break the 1Gb barrier

- Exponential increase in software complexity
- In some areas code size is doubling every 9 months [ST Microelectronics, Medea Workshop, Fall 2003]
- ... > 70% of the development cost for complex systems such as automotive electronics and communication systems are due to software development [A. Sangiovanni-Vincentelli, 1999]



Rob van Ommering, COPA Tutorial, as cited by: Gert Müller: Opportunities and challenges in embedded systems, Eindhoven Embedded Systems Institute, 2004

Course Overview

- To get an insight into the broad area of system safety
- We cover techniques for high availability, fault tolerance, monitoring, detection, diagnosis, and confinement of failure, ways to improve availability through fast recovery and graceful service degradation, and techniques for using redundancy and replication.
- We also discuss the utopia of flawless software, the impact of scale on availability, ways to cope with human operator error, and metrics for evaluating dependability.

21

Contents

- Fault tolerance
- System reliability
- Hardware redundancy
- Error detection techniques
- Coding techniques
- Processor-level detection and recovery
- Disk arrays
- Checkpointing and recovery
- Software fault tolerance
- Testing distributed real-time systems
- ...

22

Lecture Outline



- ✓ **Historical perspective and famous incidents/accidents**

- **Basic terminology**

23

Murphy's Law

- "If something can go wrong, it will go wrong"

*Major Edward A. Murphy, Jr.
US Air Force, 1949*

- "Every component than can be installed backward, eventually will be"

24

Genesis Space Capsule

- \$260 million Genesis capsule was collecting samples of the solar wind over 3 years period
- Crashed in Sept 2004 due to the failure of the parachutes
- Reason: the deceleration sensors — the accelerometers — were all installed backwards. The craft's autopilot never got a clue that it had hit an atmosphere and that hard ground was just ahead.



25

Mars Orbiter

- One of the Mars Orbiter probes crashed into the planet in 1999.
- It did turn out that engineers who built the Mars Climate Orbiter had provided a data table in "pound-force" rather than newtons, the metric measure of force.
- NASA flight controllers at the Jet Propulsion Laboratory in Pasadena, Calif., had used the faulty table for their navigation calculations during the long coast from Earth to Mars.

26

Lockheed Martin Titan 4

- In 1998, a LockMart Titan 4 booster carrying a \$1 billion LockMart Vortex-class spy satellite pitched sideways and exploded 40 seconds after liftoff from Cape Canaveral, Fla.
- Reason: frayed wiring that apparently had not been inspected. The guidance systems were without power for a fraction of a second.



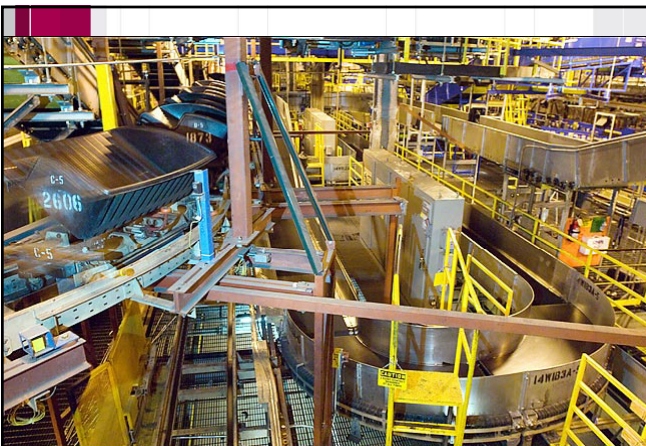
A Titan 4 rocket explodes shortly after takeoff in August 1998.

27

Therac-25

- Therac-25:
 - the most serious computer-related accidents to date (at least nonmilitary and admitted)
 - machine for radiation therapy (treating cancer)
 - between June 1985 and January 1987 (at least) six patients received severe overdoses (two died shortly afterward, two might have died but died because of cancer, the other two had permanent disabilities)
 - scanning magnets are used to spread the beam and vary the beam energy
 - dual-mode: electron beams for surface tumors, X-ray for deep tumors

28



29

Denver Airport

- Denver International Airport, Colorado: intelligent luggage transportation system with 4000 "Telecars", 35km rails, controlled by a network of 100 computers with 5000 sensors, 400 radio antennas, and 56 barcode readers. Price: \$186 million (BAE Automated Systems).
- Due to SW problems about one year delay which costs \$1.1 million per day (1993).
- Abandoned in 2005 to save \$1 million per month on maintenance

30

Lecture Outline



- ✓ **Historical perspective and famous incidents/accidents**

- **Basic terminology**

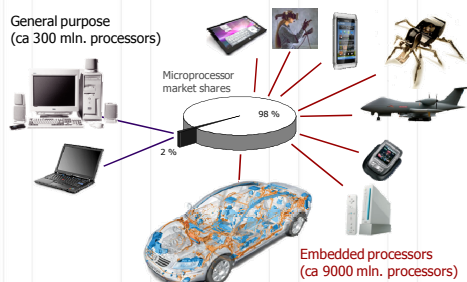
31

Embedded Systems

- Computing systems are everywhere
- Most of us think of "desktop" computers
 - PC's
 - Laptops
 - Mainframes
 - Servers
- But there's another type of computing system
 - Far more common...

32

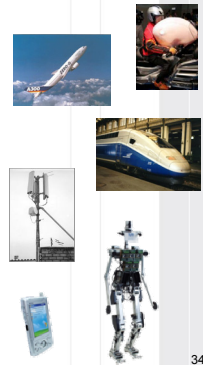
General-Purpose vs. Embedded



33

Embedded Systems, cont.

- Embedded computing systems
 - Computing systems embedded within electronic devices
 - Hard to define. Nearly any computing system other than a desktop computer
 - Billions of units produced yearly, versus millions of desktop units
 - Perhaps 50 per household and per automobile

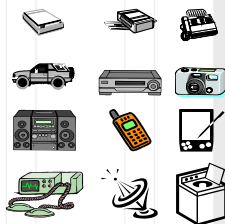


34

A "Short List" of Embedded Systems

Anti-lock brakes
Auto-focus cameras
Automatic teller machines
Automatic toll systems
Automatic transmission
Avionic systems
Battery chargers
Camcorders
Cell phones
Cell-phone base stations
Cordless phones
Cruise control
Curbside check-in systems
Digital cameras
Disk drives
Electronic card readers
Electronic instruments
Electronic toys/games
Factory control
Fax machines
Fingerprint identifiers
Home security systems
Life-support systems
Medical testing systems

Modems
MPEG decoders
Network cards
Network switches/routers
On-board navigation
Pagers
Photocopiers
Point-of-sale systems
Portable video games
Printers
Satellite phones
Scanners
Smart ovens/dishwashers
Speech recognizers
Stereo systems
Teleconferencing systems
Televisions
Temperature controllers
Theft tracking systems
TV set-top boxes
VCR's, DVD players
Video game consoles
Video phones
Washers and dryers

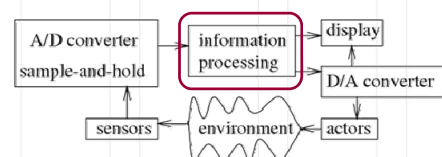


Our ~~only~~ lives depend on embedded systems

35

What is an Embedded System?

- Definition
 - an **embedded system** special-purpose computer system, part of a larger system which it controls.
- Notes
 - A computer is used in such devices primarily as a means to simplify the system design and to provide flexibility.
 - Often the user of the device is not even aware that a computer is present.



36

Characteristics of Embedded Systems

- Single-functioned
 - Dedicated to perform a single function
- Complex functionality
 - Often have to run sophisticated algorithms or multiple algorithms.
 - Cell phone, laser printer.
- Tightly-constrained
 - Low cost, low power, small, fast, etc.
- Reactive and real-time
 - Continually reacts to changes in the system's environment
 - Must compute certain results in real-time without delay
- Safety-critical
 - Must not endanger human life and the environment

37

Real-Time Systems

- **Time**
 - The correctness of the system behavior depends not only on the logical results of the computations, but also on the *time* at which these results are produced.
- **Real**
 - The reaction to the outside events must occur *during* their evolution. The system time must be measured using the same time scale used for measuring the time in the controlled environment.

38

Hard vs. Soft Real-Time

- Definitions
 - A real-time task is said to be **hard** if missing its deadline may cause catastrophic consequences on the environment under control.
 - A real-time task is said to be **soft** if meeting its deadline is desirable for performance reasons, but missing its deadline does not cause serious damage to the environment and does not jeopardize correct system behaviour.
- Definition
 - A real-time system that is able to handle hard real-time tasks is called a **hard real-time system**.

39

Hard vs. soft, cont.

- Examples of hard activities
 - Sensory data acquisition
 - Detection of critical conditions
 - Actuator serving
 - Low-level control of critical system components
 - Planning sensory-motor actions that tightly interact with the environment
- Examples of soft activities
 - The command interpreter of the user interface
 - Handling input data from the keyboard
 - Displaying messages on the screen
 - Representation of system state variables
 - Graphical activities
 - Saving report data

40

Functional vs. Non-Functional Requirements

- Functional requirements
 - output as a function of input
- Non-functional requirements:
 - Time required to compute output
 - Reliability, availability, integrity, maintainability, dependability
 - Size, weight, power consumption, etc.

41

Fault Tolerance

- A fault-tolerant system is one that can continue to correctly perform its specified tasks in the presence of failures:
 - hardware
 - software
 - user errors
 - environmental, input, ...
- Fault tolerance is the attribute that enables a system to achieve fault tolerant operation.

42



Basic Concepts

- *Fault Tolerance* is closely related to the notion of "Dependability". This is characterized under a number of headings:
 - *Availability* – the system is ready to be used immediately.
 - *Reliability* – the system can run continuously without failure.
 - *Safety* – if a system fails, nothing catastrophic will happen.
 - *Maintainability* – when a system fails, it can be repaired easily and quickly (and, sometimes, without its users noticing the failure).

43



Faults, Errors & Failures

- Fault: a defect within the system or a situation that can lead to the failure
- Error: manifestation of the fault – an unexpected behavior
- Failure: system not performing its intended function

Fault → Error → Failure

44



Fault Examples

- Bit flips in hardware due to cosmic radiation
 - A person on an airplane over the Atlantic at 35,000 ft working on a laptop with 256 Mbytes (2 Gbits) of memory. At this altitude, the SER of 600 FITs per megabit becomes 100,000 FITs per megabit, resulting in a potential error every five hours.
 - 1 FIT (failures in time), is the number of failures in 1 billion device-operation hours. A measurement of 1000 FITs corresponds to a MTTF (mean time to failure) of approximately 114 years.

45



Fault Examples

- Year 2000 bug
- Loose wire
- Aircraft retracting its landing gear while on ground
- Effects in time:
 - Permanent
 - Transient
 - Intermittent



46



Permanent

- A permanent fault or failure is one which is stable and continuous.
- Permanent hardware failures require some component to be replaced or repaired.
- An example of a permanent fault would be a VLSI chip with a manufacturing defect, causing one input pin to be stuck high (stuck-at-1).

47



Transient

- A transient fault is one which results from a temporary environmental condition.
- For example, a voltage spike might cause a sensor to report an incorrect value for a few milliseconds before reporting correctly.

48

Transient faults

- Happen for a short time
- **Corruptions of data, miscalculation in logic**
- Do not cause a permanent damage of circuits
- Causes are outside system boundaries

Electromagnetic interference (EMI)

Radiation



Lightning storms

49

Intermittent

- An intermittent fault is one which only manifests occasionally, due to unstable hardware or certain system states.
- A loose contact on a connector will often cause an intermittent fault.
- Intermittent electrical faults, as a rule, are notoriously difficult to detect. Typically, whenever the fault doctor shows up, the system works fine.

50

Intermittent faults

Manifest similar as transient faults

- Happen repeatedly
- Causes are inside system boundaries

Internal EMI



Power supply fluctuations



Crosstalk

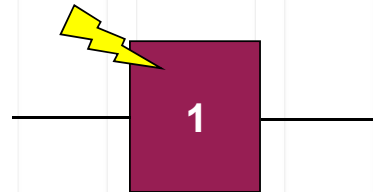


Init (Data)

Software errors (Heisenbugs)

51

Soft Errors



- Transient bit-flip (soft memory error)
 - Random event
 - Corrupts the value but not the cell
 - Can be corrected (in contrast to hard errors caused by faults in the hardware itself)
 - Happen continuously during system lifetime (i.e., can not be screened by burn-in tests)

52

Sources

- First traced to alpha particle emissions from chip packaging materials
 - Most sources removed (pure materials, different designs, shielding)
- Today's main problem: cosmic radiation
 - Cosmic particles from deep space (actually 5th- or 6th-hand collision particles)
 - At ground level ca 95% neutrons, 5% protons
 - Radioactive material in manufacturing process

53

Sources (cont.)

- Four main sources:
 - Low-energy alpha particles
 - High-energy cosmic particles
 - Thermal neutrons
 - Poor system design

SER type	Source	Mechanism	Trend
Alpha	Thorium and uranium contamination in-mold compound, silicon, or lead bumps	2- to 9-MeV alpha particle creating electron-hole tunnel traveling 25 microns in silicon	Exponential increase with scaling
Cosmic	Intergalactic sources modulated by solar flares	High-energy neutrons/protons (10 MeV to 1 GeV) colliding with silicon nuclei	Decrease in failures in time per megabit
Thermal neutron	Boron present in BPSG25-meV neutrons	Collision with B10 in BPSG	Highest, always dominates if present

54



Failures (cont.)

- **Timing** Failure: A server's response lies outside the specified time interval.
- **Response** Failure: The server's response is incorrect (value of the response is wrong, server deviates from the correct flow of control).
- **Arbitrary** Failure: A server may produce arbitrary responses at arbitrary times.

61



Fault Handling

- Fault avoidance: eliminate problem sources
 - Remove defects: Testing and debugging
 - Robust design: reduce probability of defects
 - Minimize environmental stress: Radiation shielding etc
- **Impossible to avoid faults completely**
- Fault tolerance: add redundancy to mask effect
 - Additional resources needed (more later)
 - Examples:
 - Error correction coding, voting and masking, checksums, ...
 - Backup storage, replication, ...
 - Spare tire, etc

62



Fault Tolerance

- **Fault detection** is the process of recognizing that a fault has occurred. Fault detection is often required before any recovery procedure can be initiated. The techniques include error detection codes, self-checking/failsafe logic, watchdog timers, and others.
- **Fault location** is the process of determining where a fault has occurred so that an appropriate recovery can be initiated.

63



Fault Tolerance (cont.)

- **Fault containment** is the process of isolating a fault and preventing the effects of that fault from propagating throughout the system.
- **Fault recovery** is the process of remaining operational or regaining operational status via reconfiguration even in the presence of faults. A few basic approaches are fault masking, retry, and rollback.

64



Definitions

- Failure rate (λ):
 - Average frequency with which something fails.

$$\frac{6 \text{ failures}}{7502 \text{ hrs}} = 0.0007998 \text{ failures / hr} = 799.8 \times 10^{-6} \text{ failures / hr}$$

- Mean time to failure (MTTF):
 - Average time between failures

$$MTTF = \frac{1}{\lambda}$$

65

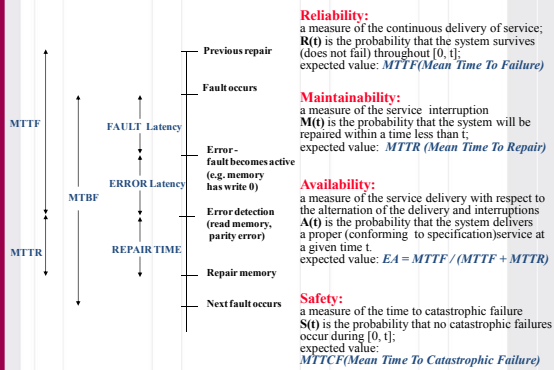


Dependability

- Property of a computing system which allows reliance to be justifiably placed on the service it delivers
- Dependability = reliability + availability + safety + security + ...
- Reliability → continuity of correct service
- Availability → readiness of usage
- Safety → no catastrophic consequences
- Security → prevention of unauthorized access

66

Dependability Concepts

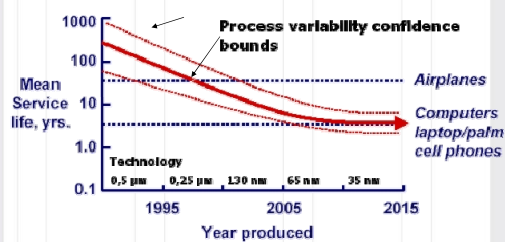


Reliability

- A measure of an it performing its intended function satisfactorily for a prescribed time and under given environment conditions.
- Probability that system will survive to time t
 - In aerospace industry the requirement is that failure probability is 10^{-9} (one failure over 10^9 hours (114 000 years) of operation)
- Time To Failure (TTF)
- Mean Time To Failure (MTTF)

68

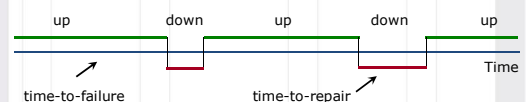
Commercial Chip Reliability Estimation



*Extrapolated from ITRS roadmap, MRQW-2002, Bernstein

69

Availability



$$Availability = \frac{MTTF}{MTTF + MTTR}$$

Availability:

- Probability that system is operational at time t

High availability:

- $MTTF \rightarrow \text{infinity}$ (high reliability)
- $MTTR \rightarrow \text{zero}$ (fast recovery)

70

Maintainability

- $M(t)$ is the probability that a failed system will be restored within a specified period of time t .
- Restoration process:
 - locating problem, e.g. via diagnostics
 - physically repairing system
 - bringing system back to its operational condition

71

Graceful Degradation

- The ability of system to automatically decrease its level of performance to compensate for hardware failure and software errors.

72

The Myth of the Nines

Nines	Availability	Downtime per year	Downtime per week	Example
2 nines	99%	3.65 days	1.7 hours	General web site
3 nines	99.9%	8.75 hours	10.1 min	E-commerce site
4 nines	99.99%	52.5 min	1.0 min	Enterprise mail server
5 nines	99.999%	5.25 min	6.0 s	Telephone system
6 nines	99.9999%	31.5 s	0.6 s	Carrier-grade network switch

73

Historical Evaluation

- Mean Time Between Failures:

$$MTBF = MTTR + MTTF$$

- ENIAC. MTBF: 7 minutes (18000 vacuum tubes)
 - ENIAC → TX-2 interactive computer (MIT) → web
- F-8 Crusader – first fly-by-wire, 375 hours → 750 hours (IBM AP-101)
 - MD-11
 - A320 family
- Patriot missile defence system
 - 1/3 sec in 100 hours, targeting error: 600 m
 - Needed reboot after 8 hours, was learned in hard way...

74

Ultra-Reliable Systems

- Airbus A320 family fly-by-wire system:
 - computer controls all actuators
 - no control rods, cables in the middle
 - 7 central flight control computers
 - 3 Motorola 68000
 - 2 Intel 80C86
 - 2 Intel 80C286
 - software for hardware written by different software houses (C, ASM, dedicated one, specifically developed)
 - all error checking & debugging performed separately
 - computer allows pilot to fly craft up to certain limits (flight envelope)
 - beyond: computer takes over

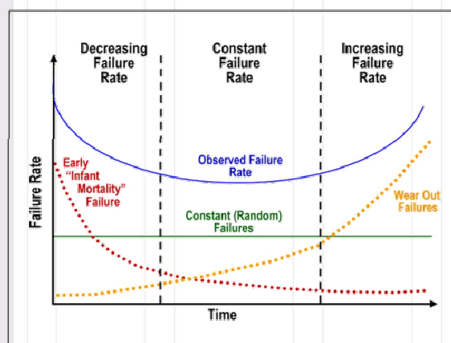
75

Hardware and Environment Failures

- Moving parts, high speed, low tolerance, high complexity: disks, tape drives/libraries
- Lowest MTBF found in fans and power supplies
- Often fans fail gradually → subtle, sporadic failures in CPU, memory, backplane
- Environment: power, cooling, dehumidifying, cables, fire, collapsing racks, ventilation, earthquakes, ...

76

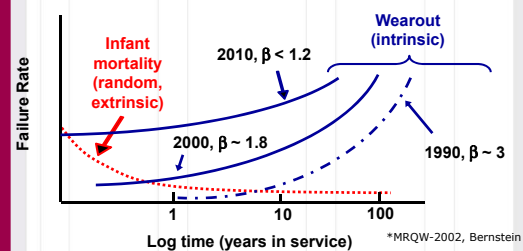
Bathtub Curve



77

Device Reliability Trends

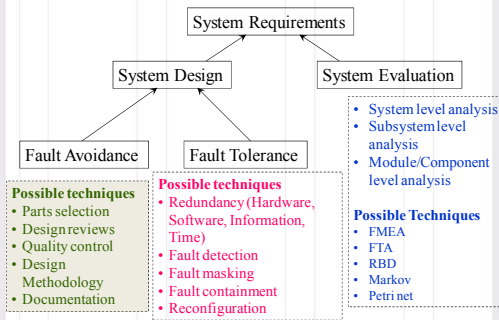
- As technology progresses, wearout failures become statistically indistinguishable from infant mortality failures with the same wearout drivers.



*MRQW-2002, Bernstein

78

System Design & Evaluation Top-Level View



79

Safety

- Attribute of a system which either operates correctly or fails in a safe manner
- Freedom from exposure to danger, or exemption from hurt, injury or loss.
- “Fail-safe”: traffic lights start to blink yellow
- Degrees of safety
- Closely related to risk

80

Risk

- A combination of the likelihood of an accident and the severity of the potential consequences
- The harm that can result if a threat is actualised

- Acceptable/tolerable risk: The Ford Pinto case (1968)

BENEFITS

Savings: 180 burn deaths, 180 serious burn injuries, 2,100 burned vehicles.
 Unit Cost: \$200,000 per death, \$67,000 per injury, \$700 per vehicle.
 Total Benefit: $180 \times (\$200,000) + 180 \times (\$67,000) + 2,100 \times (\$700) = \$49.5 \text{ million}$.

COSTS

Sales: 11 million cars, 1.5 million light trucks.
 Unit Cost: \$11 per car, \$11 per truck.
 Total Cost: $11,000,000 \times (\$11) + 1,500,000 \times (\$11) = \$137 \text{ million}$.

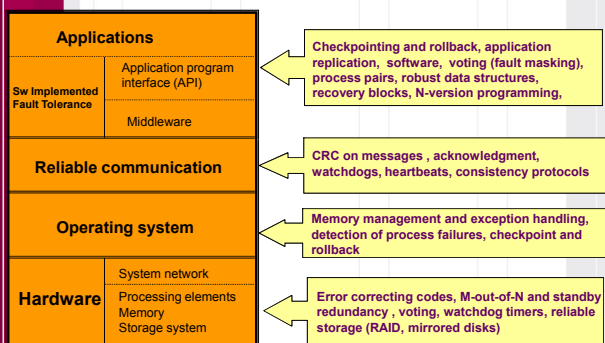
81

System Safety & Hazards

- Safety:
 - achieved by anticipating accidents and eliminating their causes
- Hazards are potential causes of accidents
 - Conditions in a system which together with other factors in the environment inevitably cause accidents

82

Reliability is a System Issue



[9er]

Questions?

Gert Jervan



Administrative issues

www.pld.ttu.ee/IAF0530

Gert Jervan
IT-229 620 2261
gert.jervan@pld.ttu.ee
www.pld.ttu.ee/~gerje

- Case Studies
 - Presentation + report
- Exam

© Gert Jervan

85