Designing reliable systems with unreliable components

PRADIP BOSE
pbose@us.ibm.com

The problem formulated in the title above is an old one; it has been with us in the electronics industry from the very beginning. In the days of vacuum tubes, component-level reliability was never a strong point, and the possibility of an unacceptably small mean-time-to-failure (MTTF) for the target system was one of the biggest risks design teams took.

Consider, for example, the pioneering ENIAC machine, which had 17,468 vacuum tubes, 7,200 crystal diodes, 1,500 relays, 70,000 resistors, 10,000 capacitors, and about 5 million hand-soldered joints (see http://en.wikipedia.org/wiki/ENIAC). Many electronics experts predicted that component failures (in particular, tube failures) would be so frequent that the machine would never be useful. In the initial years following ENIAC’s 1947 commission for continuous operation, this prediction appeared at least partially correct. Several tubes burned out almost every day, leaving it nonfunctional about half the time.

However, the engineers (system architects) and component manufacturers improved their art over time to improve the system’s availability. By 1948, special high-reliability tubes were available. In addition, engineers made an important observation from the log book of system checkstops: Most failures occurred during the warm-up and cool-down periods, when the tube heaters and cathodes were under the most thermal stress. By the simple (if expensive) expedient of never turning the machine off, the engineers reduced ENIAC’s tube failures to the more acceptable rate of one every two days. In addition, fault diagnosis, location, and repair were evidently quite advanced; the average down time was just a few minutes.

According to a 1989 interview with J. Presper Eckert, one of ENIAC’s designers, good engineering converted the story of the continuously failing tubes to a myth: “We had a tube fail about every two days, and we could locate the problem within 15 minutes.” In the field of reliability engineering, failure rate is often measured in units of FIT (failures in time); 1 FIT means one failure in a billion hours. The system MTTF is computed in hours as \(10^9/(\text{system FITs})\). For ENIAC, ignoring the failure rate of the other components for the moment (since the FIT rate contribution from the 17,468 tubes far dominated the overall FITs), the 48-hour MTTF translates roughly to a per-tube effective FIT rate of 1,192. This is a remarkably small number, since it inverts to a per-tube MTTF of about 95 years! In words, the effective mean time to failure of a single component tube (with its associated peripheral circuitry) was 95 years: much, much longer than a single vacuum tube’s specified stand-alone, component-level reliability. In fact, in 1954, ENIAC’s longest continuous period of operation without a failure was 116 hours—close to five days.

Such remarkably low failure rates stand as a tribute to the precise engineering of ENIAC. They should also serve as an inspiration to chip- and system-level designers today. The problems of deep-submicron technologies increasingly highlight component-level unreliability, but now the components are chips (or cores, interconnect, and storage subunits within a chip). The largest supercomputers being planned today have thousands (or even hundreds of thousands) of such components. But the transistor-level FITs are minuscule compared to the tube-level FITs of yesteryear; the effective chip-level FITs are on the order of those measured for the tubes of the ENIAC age.

Although this progress is truly remarkable, the sheer growth in the number of components (on a chip and within a full system), fueled by advances in technology, continues to challenge architects and design engineers. Moreover, the threat of diminished component reliability—stemming from escalating thermal profiles, process-level variability, inductive noise on supply voltage rails, soft error rates, and so on—offers further challenges for future chip and system designers.

Three articles in this general issue of IEEE Micro address the challenge of reliable designs of the future: Neto et al. write about detecting soft errors via built-in current sensors. The article by Nepal et al. proposes a solution for increasing the chip-level immunity to single-event upsets and noise. Teodorescu et al. describe an efficient cache-level checkpointing and rollback scheme.

The other three articles do not address reliability directly, but describe solutions to...
problems that relate to reliable (robust) system performance. The article by Gonzalez describes an interesting approach to providing extensible and reconfigurable instruction set architectures, to provide robust, predictable performance across a diverse set of applications. The article by Garcia et al. deals with the problem of congestion management for interconnection networks. A good solution here ensures robust, scalable performance in multiprocessor systems. In the article on Blue Gene/L by Salapura et al., we find an intelligent exploitation of thread- and data-level parallelism to deliver scalable, large-scale performance at affordable power. Power and thermal efficiency of such large supercomputing systems are prerequisites for ensuring tolerable MTTF values, so this article addresses a fundamental issue that ultimately also leads to the delivery of a reliable system.

Pradip Bose
Editor in Chief
IEEE Micro