Reliability, Availability, and Serviceability

Lecture #16: Tuesday, 22 May 2000
Lecturer: David Lie
Scribe: John Maly

1 Overview of RAS

Definition: RAS

RAS stands for Reliability, Availability, and Serviceability. It actually encompasses much more than just the processor of a system, but this discussion will primarily be confined to the processor-related aspects of RAS (although some discussion of redundant power supply, cooling fans, and interconnect does take place at the end).

Why is RAS important?

Reliability is crucial in modern systems, even if it comes at the expense of system performance. Slowness can be an acceptable trait of a system, but failure and data loss are almost never acceptable. Downtime is equally unacceptable, lending to the obvious importance of availability. Finally, serviceability contributes to both of the aforementioned traits, and should help to reduce the ongoing cost of running the system. Some definitions are given below.

Definition: Reliability

Reliability is a function of time that expresses the probability at time t+1 that a system is still working, given that it was working at time t.

Definition: Availability

Availability is the measure of how often the system is available for use (such as a system's up-time percentage). Availability and reliability may sound like the same thing, but it is worth noting that a system can have great Availability but no Reliability. An internet router is a good example of this; it stores no state data. It is one of the few systems wherein data loss is acceptable, as long as high availability is maintained.

Definition: Serviceability

Serviceability is a broad definition describing how easily serviced or repaired a system is. For example, a system with modular, hot- swapable components would have a good level of serviceability.

Note that implementing hot-swapable components contributes to all three of the above qualities, not just serviceability.

Some Trends Currently Effecting RAS in Processors:
1. Shrinking achitectures and capacitances
2. Increased processor complexity (difficult to fully verify)
3. Extensive use of dynamic logic (more discharging)
(The latter is an issue with static logic also, but dynamic logic has lower margins for error)

What are the two major themes in these papers for implementing RAS?

Data integrity: The ability to isolate failures is critical. We want to have safe failures (also known as fail-stop).

Redundancy: Redundancy in functional units, interconnects, and backup processors is an effective way to assure reliability and availability, and to prevent the customer from needing to service the product.

What are the three steps to failure recovery?
1. Detect the failure
2. Isolate it from causing further problems
3. Fix the failure and any associated side effects

2 DIVA: A Reliable Substrate for Deep Submicron Microarchitecture Design

How does DIVA work?

DIVA uses check-compute units. After each instruction, these units are fed the end result. They then compute the instruction result a second time, and compare their result

to the supplied (initial) result. If the results do not match, the processor goes back and performs the instruction again.

DIVA was designed with simplicity in mind. To this end, the checker is always assumed correct (no voting takes place). This helps to reduce the design fault possibility. DIVA is an improvement on Tandem, because it will actually catch design errors (which would, theoretically, produce agreeing, incorrect results).

Definitions: Validation vs. Verification

While often used interchangably, these terms mean different things. Validation refers to making sure a design functions according to specification. This can be done with a simulator in software. Conversely, verification refers to making sure a finished product (e.g., a processor) operates as intended (or previously simulated).

What does a Watchdog Timer do?

A watchdog timer exists to catch deadlock situations and restart things if they occur (when the timer reaches a certain high value). This guarantees that the processor core makes forward progress. This timer gets reset when the process it is timing is retired.

What are the main sources of performance impact in DIVA?
1. Number of memory ports (extra ports yield better performance)
2. Latency of detector (if this is really long, processes can't retire quickly, and the instruction window fills up)
3. Varied exception rates (higher numbers of exceptions mean that less real work is getting done)

Final note: The in-order core has almost no dependency. Ideally, it has 100 percent prediction because it runs "in the wake" of the superscalar processor. There is no forwarding (except for the comm-checking bypass; we cannot perform checks without the correct value here).

3 Ultra Enterprise 10000 Server: SunTrust Reliability, Availability, and Serviceability

Features:
1. Interconnect redundancy (uses crossbar interconnect)

2. System service (a separate processor monitors the whole system looking for and correcting errors
3. Dynamic configuration (hotswap ability and corresponding software support)

Note that hotswapping entails a fairly complex procedure. The system must first evict data, then evict each process after its next context switch, and finally shut down power to the component before it can be removed. Processes are migrated over to other components and execution is resumed.

Fault-Tolerant Cooling and Power

This system has redundant power supplies and power lines. The power supplies overlap (current comes from all of them simultaneously). The system uses "n+1 redundancy", meaning that one power supply can fail and the system continues to function, as it needs only n working supplies at a given time.

The system also has redundant cooling fans. When one fan dies, the remaining fans are actually sped up to compensate for the loss of airflow!

Partitioning

Partitioning was, prior to this machine, only available in mainframes. Essentially, the system is divided into domains. Processes are distributed to the various domains, and redistributed when failure necessitates it. The system uses a crossbar interconnect to facilitate this redistribution.

Automatic System Reconfiguration

The Enterprise 10000 shares peripherals, IO, and the backplane. Memory is not contiguous, as it is spread across different boards. The backplane is the only device that is not hot-swapable.

Design note: It is generally inadvisable to place active devices onto backplanes when designing a system. Such devices cannot be replaced without replacing the backplane, which in most cases means the machine must come down for prolonged service.

4 IBM's S/390 G5 Microprocessor Design

What makes this processor reliable?

IBM claims that this processor is 100 percent effective in correcting soft errors. This reliability comes from several features:
1. Multiple processors in 2 regions (processes can swap out if one processor goes down); 12 real processors, 12 "spare". (These are, however, in a multi-chip module, so a single defective processor cannot be replaced. It is assumed that the 12 extra processors will be enough redundancy for the life of the system.)
2. Implemented error correction and recovery (see below)
3. Implemented array recovery for hard errors (see below)
4. Uses CMOS instead of bipolar transistors
5. Uses decimal ALU's instead of binary for more precision
6. 20 percent of all logic in the processor is devoted to error-checking

Hardware error recovery takes place in a series of steps:
a. Stop the system upon error detection
b. Any queued data is written into the L2 cache
c. Most of the machine's functional units are reset
d. Use error-correcting codes:
If correct, continue.
If not, update the register file with the error-correcting code (ECC)
result, and assume that it is the correct value.
e. Repeat step d (one time).

It is interesting to note that parity is used in the L1 cache, and ECC used in the L2 cache. The rationale for this is that redundant data uses only parity; if an error is detected, a new copy of the data can be brought in. If the data is unique, however, error-correcting codes are used to attempt to reconstruct/correct it.