The story of instant recovery – Extended abstract –

Goetz Graefe
Hewlett-Packard Laboratories

This presentation will summarize the history and the technology of instant recovery from system and media failures. The story starts with modern hardware, e.g., flash storage, and the danger of localized failures due to limited write endurance. Initial research sought methods for detection and recovery of localized, i.e., single-page failures.

Figure 1: Single-page failures. [GK12]

Figure 1 distinguishes multiple failure classes. If single-page recovery is available, a storage device remains useful even after a localized failure. Otherwise, a single-page failure may imply a device failure, which in turn may imply a node failure - clearly very undesirable if it is avoidable.

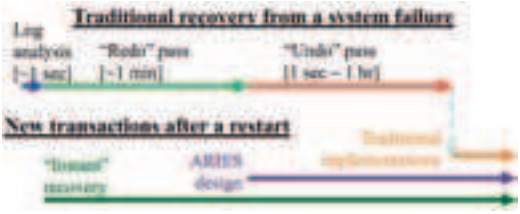

Figure 2: Instant recovery from system failures.

Figure 2 illustrates an application of single-page recovery, specifically incremental recovery from a system failure, e.g., a process crash. Traditional implementations permit new database transactions only after all recovery activities are complete, which may take

minutes or even hours. Instant recovery after a system failure requires only log analysis. Differently than advanced techniques by Mohan and by Speer and Kirchberg, which give early access only to pages not modified for a long time, instant restart permits immediate access to the pre-crash working set. Differently than all previous recovery techniques, new transactions guide the recovery, not scans of the recovery log. Figure 4 illustrates what instant recovery means for media failures. While traditional techniques enable new database transactions only after hours of recovery efforts using database backups and log archives, instant media recovery enables new database transactions almost immediately, i.e., within seconds of an empty replacement device becoming available. The core technique is on-demand incremental recovery enabled by novel log archiving. Due to a single pass over the recovery log, log archiving remains efficient.

Figure 3: Instant recovery from media failures.

Figure 4 shows how to extend these ideas to index operations without detailed logging, i.e., with allocation-only logging. Instead of detailed ”redo” information in the recovery log, controlled retention of intermediate files enables efficient single-page recovery from failures.

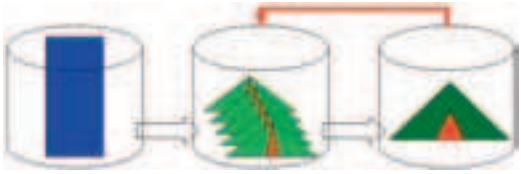

Figure 4: Instant recovery from media failures. [GG13]

The first grand prize for this research will be individual nodes with greatly improved mean time to repair and thus greatly improved availability. The second grand prize will be instant failover to a node that has an out-of-date copy of the database plus log archive and recent recovery log. The third grand prize will be a data center design in which high availability for N working nodes is possible with N+3 nodes rather than with today’s 3N nodes. In other words, we hope to reduce costs for machines, cabling, power, cooling, space, etc. to about a third of today’s standard data center design.

(Figures copied from the author’s cited papers.)

References

[CM92] B. G. Lindsay H. Pirahesh P. M. Schwarz C. Mohan, D. J. Haderle. ARIES, a transaction recovery method supporting fine-granularity locking and partial rollbacks using write-ahead logging. In *ACM TODS 17(1)*, 1992.

[GG13] B. Seeger G. Graefe. Logical recovery from single-page failures. In *BTW*, 2013.

[GK12] Goetz Graefe and Harumi A. Kuno. Definition, Detection, and Recovery of Single-Page Failures, a Fourth Class of Database Failures. *PVLDB*, 5(7):646–655, 2012.

[Gra12] G. Graefe. Instant recovery from system failures. Unpublished manuscript, 2012.

[Gra13] G. Graefe. Instant recovery from media failures. In preparation, 2013.

[JS07] M. Kirchberg J. Speer. C-ARIES, A multi-threaded version of the ARIES recovery algorithm. In *DEXA*, 2007.

[Moh93] C. Mohan. A cost-effective method for providing improved data availability during DBMS restart recovery after a failure. In *VLDB*, 1993.