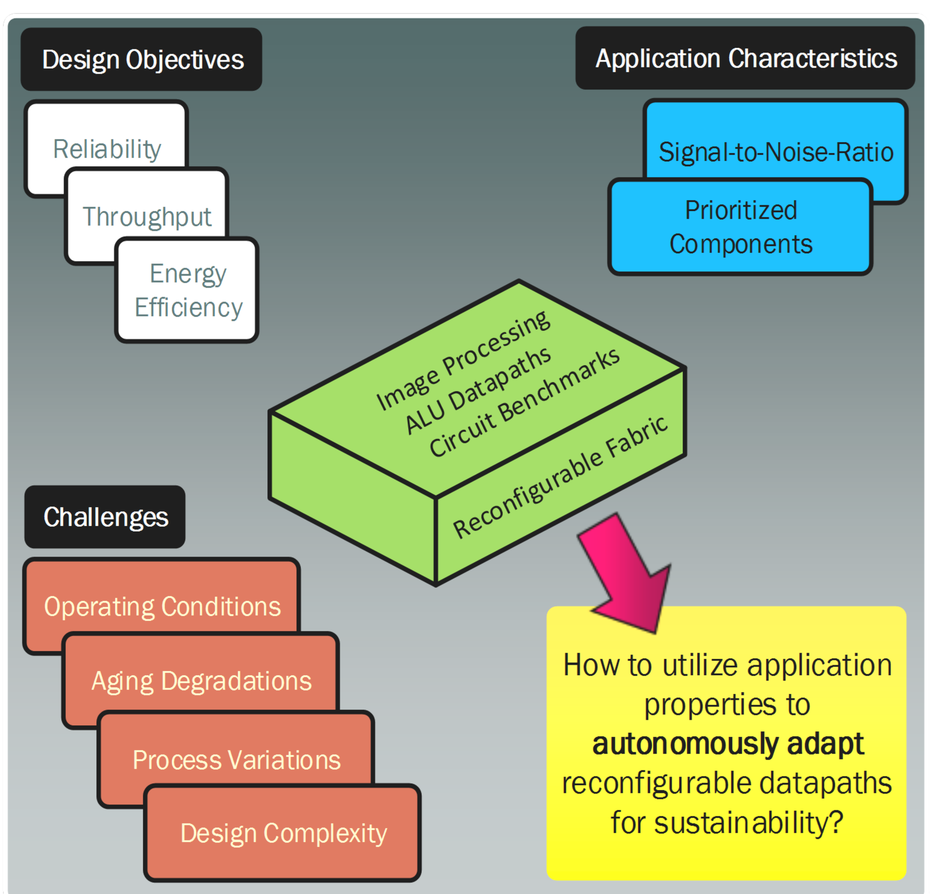
Evaluating Quality and Resilience of an Embedded Video Encoder against a Continuum of Energy Consumption

Naveed Imran, Rizwan A. Ashraf and Ronald F. DeMara
University of Central Florida

Abstract

An adaptive redundancy-based **fault-handling approach** exploiting the partial dynamic reconfiguration capability of SRAM-based FPGAs is evaluated. *Fault detection* in **Signal Processing Systems** is accomplished using a simplex hardware arrangement while a deterministic *fault isolation* scheme is employed, which neither requires test vectors nor suspends the computational throughput. The approach is validated by implementation of **Discrete Cosine Transform (DCT)** and **Motion Estimation (ME)** blocks for a H.263 video encoder benchmark in Xilinx Virtex-4 FPGA.



Introduction

- **Need for Autonomous Fault-handling** [1]
  - *unpredictable environments with limited diagnostics*
  - *technology scaling impact on reliability* [2]
- **Reconfigurable Fabrics**
  - *enable novel adaptive recovery approaches*
- **"Beyond Redundancy"** [3]
  - *overcoming constraints of design-time approaches*
- **Resource Escalation** [4]
  - *enables a continuum of energy vs. quality tradeoffs*

Research Contributions

Oblivious Fault-Detection: Intrinsic measurement of applications' health-metric using feedback loop → Simplex operation for most of mission.

Desirable Fault-Isolation: System is kept online while concurrent error detection is performed using actual runtime inputs → No need for test vectors. [5]

Degraded Quality vs. Energy Consumption: Resources computing least priority functions can be reconfigured → Throughput is application-regulated.

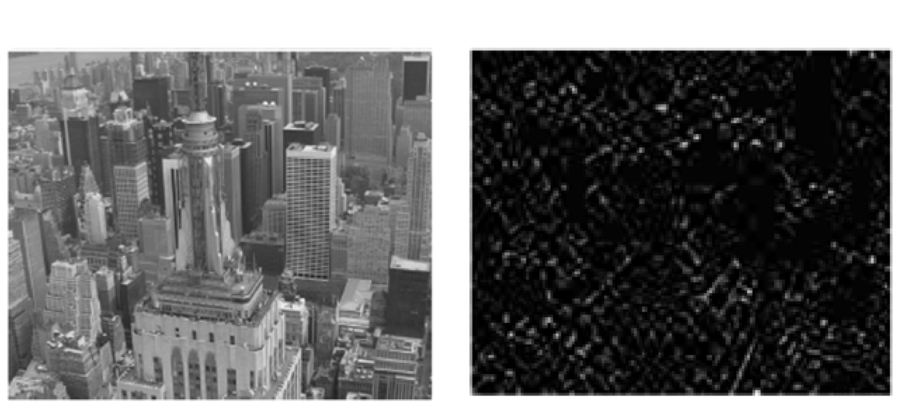


Figure 1: Health-Metric-based Fault-Handling Strategy, motivating example showing image reconstruction with (a) fully functional DCT module, PSNR=35.27dB (b) faulty PE in DCT module which computes DC-coefficient (***more significant to output quality than PEs computing AC coefficients***), PSNR=7.07dB.

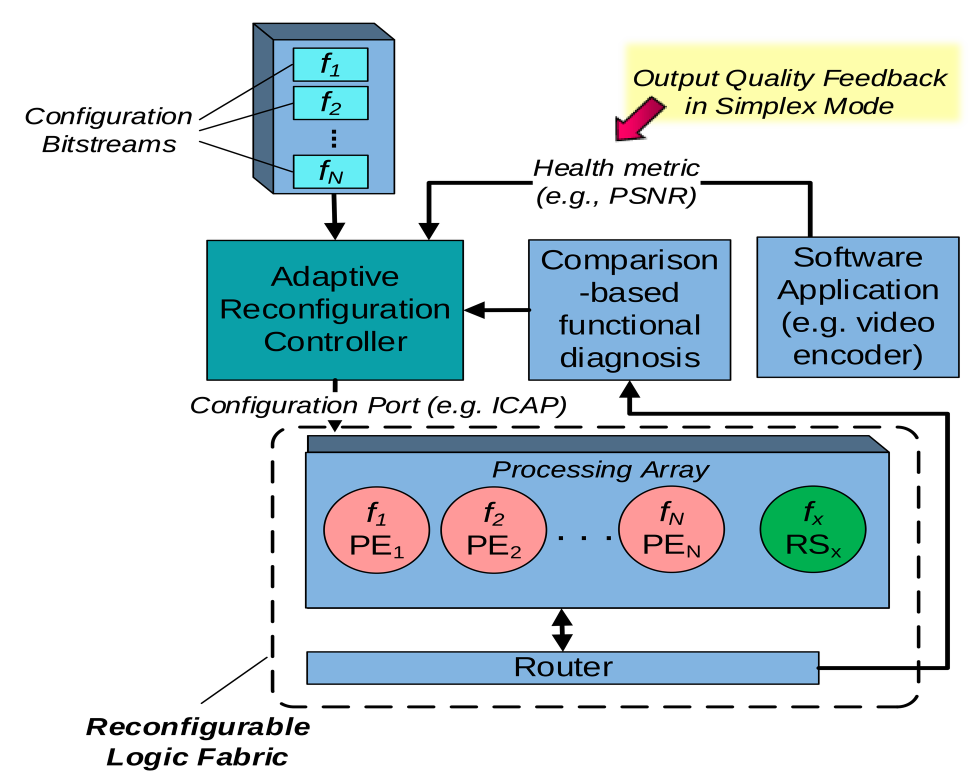


Figure 2: FPGA-based **video encoder platform**. 'PSNR' is used as a health-metric in closed loop to regulate energy/resilience/quality.

Resource Escalation

- Allocate *Reconfigurable Slack (RS)* depending upon *input signal characteristics* and area margin
- *Time-multiplex* the RS with different functions to compare their outputs with other active PEs
- Faulty PEs identified in bounded reconfigurations
- **(FaDReS) Fault Demotion using Reconfigurable Slack** [4]
  - Identified *healthy RS* is utilized to achieve diagnosis of all resources in the datapath
  - Focuses on completion of *fault isolation* phase
- **(PURE) Priority Using Resource Escalation** [6][7]
  - Identified *healthy RS* is utilized immediately for computation of highest-priority function
  - Focuses on availability and quality of throughput during *fault recovery* phase

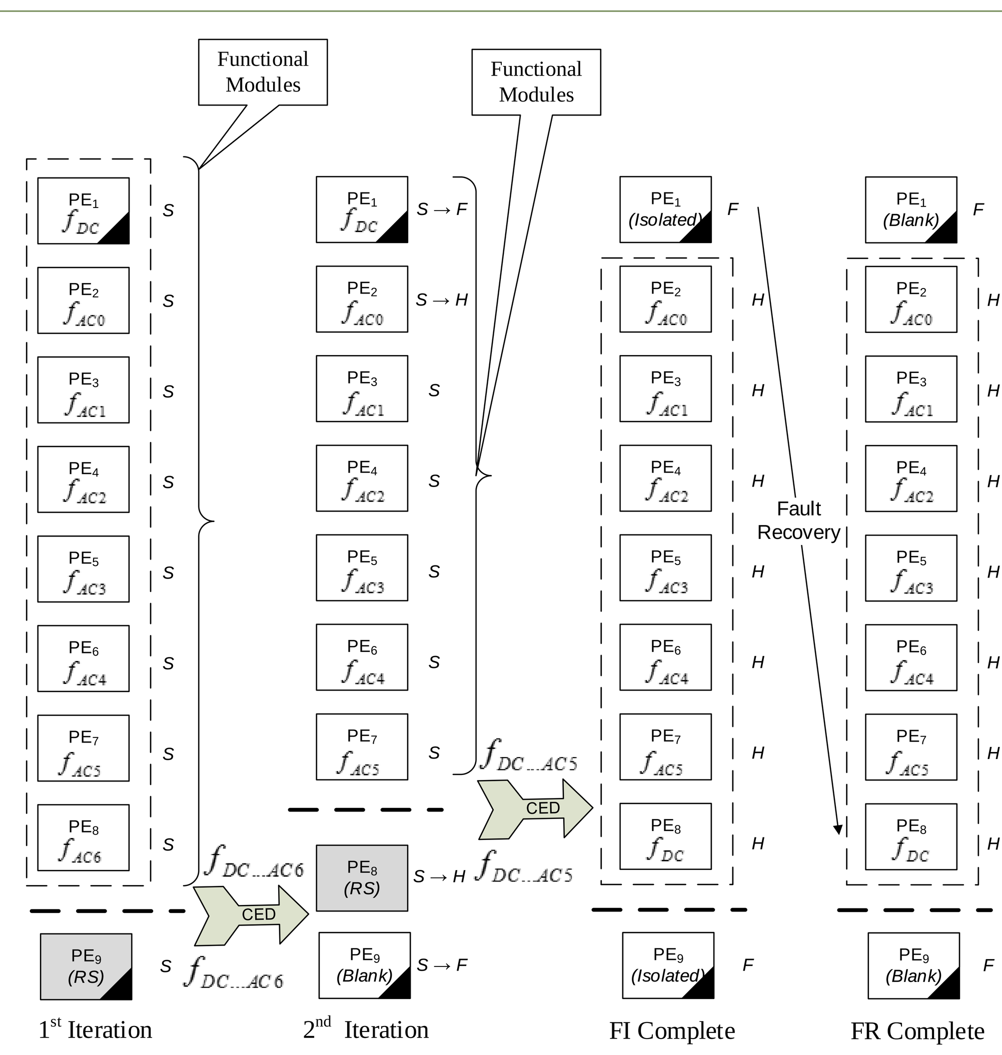
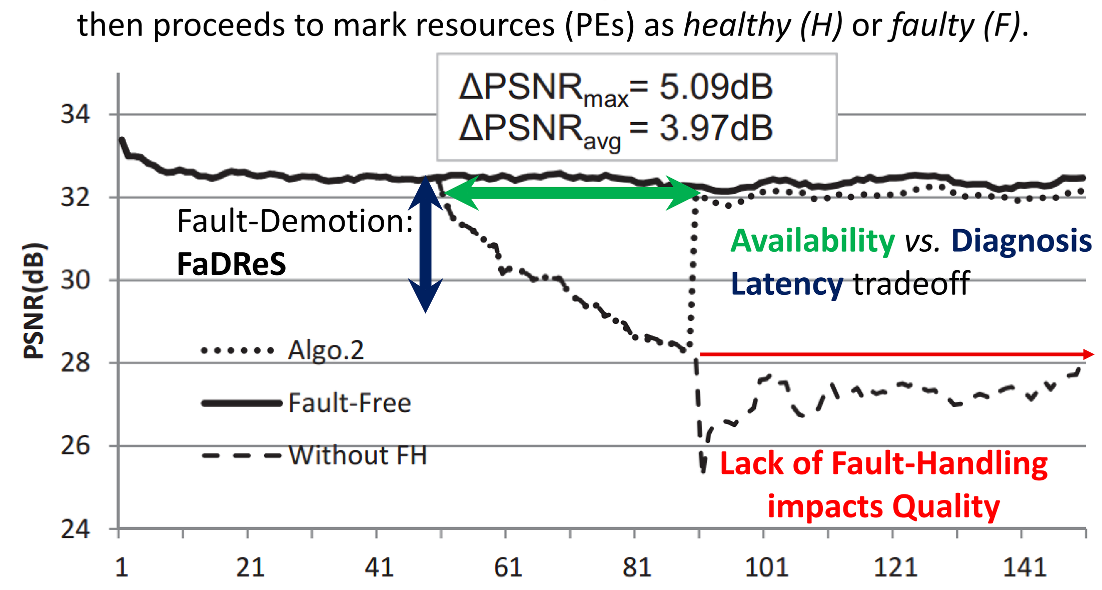


Figure 3: Fault Isolation (FI) and Recovery (FR) for **1D 8-point DCT**. Here, PE_1 (active) and PE_9 (RS) are the faulty PEs which need to be identified and removed from the datapath. Initially, all resources (PEs) are deemed *suspect (S)*. FaDReS starts by identifying a *healthy RS* (2nd iteration) and then proceeds to mark resources (PEs) as *healthy (H)* or *faulty (F)*.



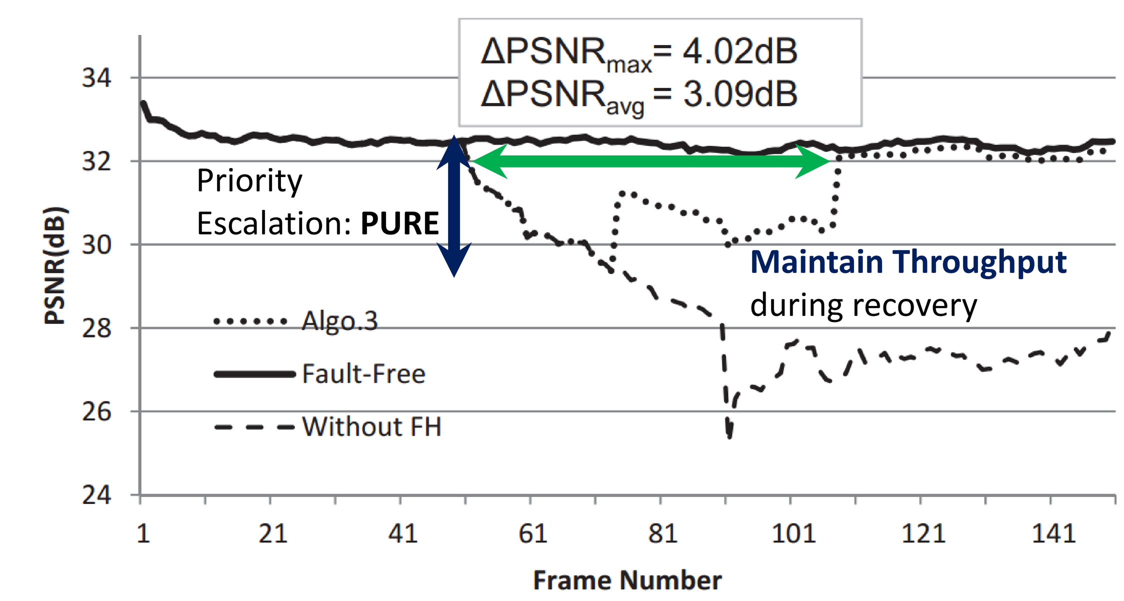


Figure 4: Fault injection and recovery in **DCT module** for **H.263 video encoder** block using different algorithms 1) FaDReS, 2) PURE. Fault-Handling Mechanism is triggered when there is a difference of 3% in PSNR (health-metric). Faults are injected in PE_1 (DC-coefficient) and PE_4

Table 1: Dynamic Energy Consumption of FaDReS during fault isolation for various fault rates in terms of number of faulty modules (N_f). Energy is calculated by product of power consumed during FI and latency of FI

Number of faulty, N_f	1	2	3	4	5	6	7
FI Latency (sec)	3.5	4.9	6.1	7.0	7.7	8.2	8.4
E (Joules)	0.91	1.27	1.57	1.81	1.99	2.16	2.18

Resource Prediction Algo [8]
- Predict S based on *Motion Vectors (MV)*
- Motion can be classified as either high, medium or low based on MVs
- Adapt resources based on predicted value of S

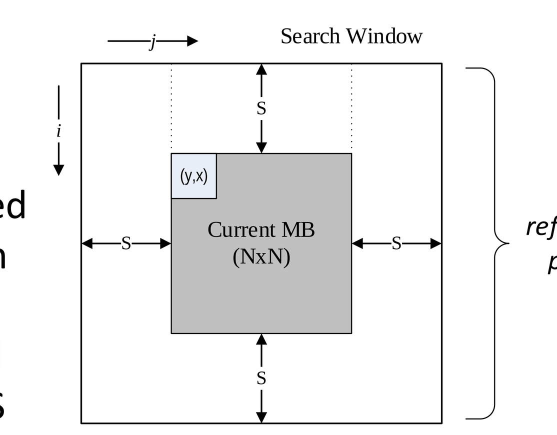


Figure 5: Fault-Handling Motion Estimation (FHME). Computation of a motion vector spatially along *j*-axis in a reference frame's search window *S* is shown. Proposed algo can adapt *S* based on type of motion.

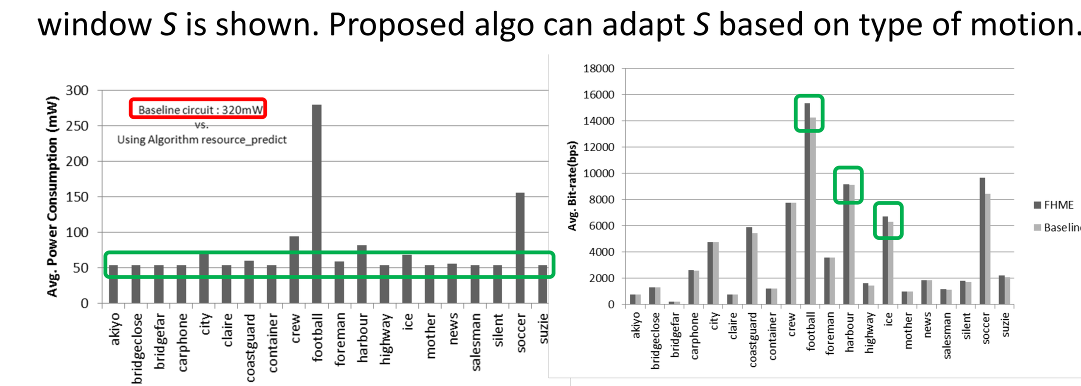


Figure 6: Motion Estimation, adapting power vs. quality using resource predict algorithm. Here, lower bit-rate implies better performance.

Video Benchmark	Motion Activity	Baseline Bit-rate	FHME Bit-rate	No. of RS
Soccer	*High*	8.43	8.62	1
Football	*High*	14.22	14.61	2
Ice	*Medium*	6.29	6.38	4
Suzie	*Low*	2.05	2.07	5

Table 2: Number of RS created using resource predict algorithm which utilizes motion information while maintaining bitrate within 3%.

Condition	Average Bit-rate (kbps)	Avg. Increase in Bit-rate
Fault-free ME	3.75	0.0% (ref)
Faulty baseline ME	8.17	117.4%
FHME (single RS)	5.25	39.7%
FHME (pair of RSs)	4.68	24.6%

Table 3: Bit-rate of encoded bitstreams for `foreman` video sequence using various architectures.

Conclusions

Energy-Aware Fault-Handling
- A simplex configuration is shown to be sufficient for applications such as DCT when a health-metric such as PSNR is available

Graceful Degradation during Diagnosis
- Degradation spanned a few frames, during which time a partial throughput is available, as an intrinsic provision of degraded mode

Priority-aware Fault Recovery
- Healthy resources are utilized for most-significant computations

References

1. X. Iturbe et. al., "R3TOS: A Novel Reliable Reconfigurable Real-Time Operating System for Highly Adaptive, Efficient, and Dependable Computing on FPGAs," *IEEE Transactions on Computers*, 2013.
2. Jörg Henkel et. al., "Reliable on-chip systems in the nano-era: lessons learnt and future trends," Design Automation Conference (DAC), 2013.
3. E.P. Kim and N.R. Shanbhag, "Soft N-Modular Redundancy," *IEEE Transactions on Computers*, 2012.
4. N. Imran, J. Lee and R. F. DeMara, "Fault Demotion Using Reconfigurable Slack (FaDReS)," *IEEE Trans. Very Large Scale Integration Systems*, 2013.
5. M.A. Wahlah and K. Goossens, "TeMNOT: A test methodology for the non-intrusive online testing of FPGA with hardwired network on chip," Microprocessors and Microsystems, 2013.
6. N. Imran, R. F. DeMara, J. Lee and J. Huang, "Self-Adapting Resource Escalation for Resilient Signal Processing Architectures," *J. of Signal Processing Systems*, 2013.
7. N. Imran, R. A. Ashraf, and R. F. DeMara, "Power and Quality-Aware Image Processing Soft-Resilience using Online Multi-Objective GAs," *IJCVR*, *in-press*, 2014.
8. N. Imran, R. A. Ashraf, J. Lee and R. F. DeMara, "Activity-Based Resource Allocation for Motion Estimation Engines," *JCSC*, *under Review*, Aug 2013.

Contact

Rizwan Ashraf
Department of Electrical Engineering and Computer Science
University of Central Florida, Orlando, FL 32816
Email: rizwan.ashraf@knights.ucf.edu
Website: cal.ucf.edu