A Dual-Mode Hybrid ARQ Scheme for Energy Efficient On-Chip Interconnects

Bo Fu and Paul Ampadu

Department of Electrical and Computer Engineering
University of Rochester, Rochester, 14627, USA
{bofu,ampadu}@ece.rochester.edu

Abstract. In this paper, we propose a dual-mode hybrid ARQ scheme for energy efficient on-chip communication, where the type of coding scheme can be dynamically selected based on different noise environments and reliability requirements. In order to reduce codec area overhead, a hardware sharing design method is implemented, resulting in only a minor increase in area costs compared to a single-mode system. For a given reliability requirement, the proposed error control scheme yields up to 35% energy improvement compared to previous solutions and up to 18% energy improvement compared to worst-case noise design method.

Key words: Adaptive error control, on-chip interconnects, hybrid ARQ, interleaving

1 Introduction

On-chip interconnect errors, exacerbated by very-deep submicron (VDSM) technology, can be caused by supply voltage fluctuation, crosstalk, process variation, radiation and electromagnetic interference [1]. Error control schemes, such as automatic repeat request (ARQ), forward error correction (FEC) and hybrid ARQ (HARQ), are widely used to improve reliability of on-chip interconnects in VDSM technology [1]-[6]. Each error control scheme has different area, power, throughput, and error correction capability trade-offs. Selection of the proper scheme can be a design-time decision based on quality-of-service (QoS) requirements [2, 3]; however, noise conditions vary with different environmental factors (e.g. temperature) and operational conditions (e.g. supply voltage), and designing for the worst case can waste energy [5]. To achieve energy efficiency, an error control scheme is needed which can intelligently provide appropriate error control capability based on noise conditions or system requirements [5, 6].

In this paper, we focus on HARQ schemes, which combine FEC and ARQ to achieve a good balance of reliability and energy consumption [1, 2]. The proposed dual-mode HARQ scheme can be dynamically configured in different noise environments. Further, in order to reduce the codec area overhead, a hardware sharing method is introduced. The proposed dual-mode HARQ method is presented in Section 2. In Section 3, the design is evaluated in terms of reliability, area and energy consumption. Conclusions are presented in Section 4.

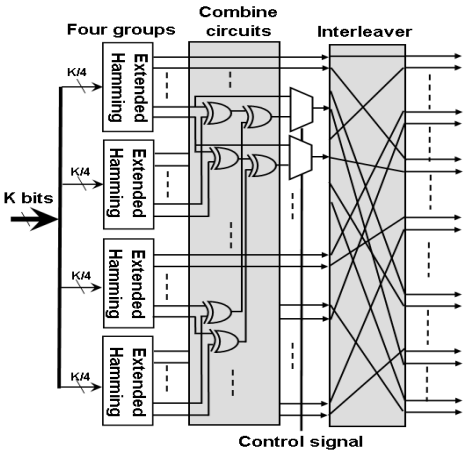

Fig. 1. Proposed dual-mode HARQ scheme.

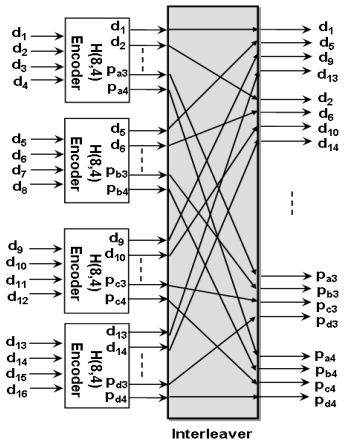

Fig. 2. Interleaving for 16-bit information.

2 Proposed Error Control Scheme

In [1, 2], single-error correcting double-error detecting (SEC-DED) codes (e.g. extended Hamming) are used to perform HARQ. As technology scales, spatial burst errors, where several adjacent bus lines are erroneous, become more common because of crosstalk effects [1]. In order to improve error resilience against burst errors, a wide bus can be split into smaller groups and encoded separately, and the outputs of these small groups can be interleaved to reduce the probability of multiple errors occurring within the same group [3]. Unfortunately, separating into groups and interleaving increase link energy consumption because of the large wire requirements.

In this paper, we propose a dual-mode HARQ scheme, which combines a traditional HARQ scheme using SEC-DED codes with interleaving. The proposed method works in two modes–(a) directly using the traditional HARQ method, or (b) separating the input message into four groups and encoding each group with a SEC-DED code, then interleaving the outputs of each group. In order to reduce the area overhead, a hardware sharing method is introduced. Fig. 1 shows the block diagram of the proposed dual-mode HARQ scheme. The input information is split into four identical groups. Each group is encoded with extended Hamming codes. The parity check bits of each group can be directly sent to the interleaver (mode b) or combined to generate the parity check bits of another SEC-DED code which uses the whole message as an input (mode a). MUXs are used to select the appropriate parity check bits for a mode based on control signals, which can be generated using the method in [5]. Fig. 2 shows an example of the interleaving relationship for a 16-bit input message.

The following example demonstrates the proposed dual-mode HARQ design. Consider a 16-bit input message, which is separated into four groups. Each group is encoded using an extended Hamming code H(8,4) with the generator matrix in Eq. (1). The parity check bits of each group can be combined to generate parity

$$G_1 = [\, I_{4\times4} \mid P_{4\times4} \,] = \left[\begin{array}{cccc|cccc} 1&0&0&0&1&1&0&1 \\ 0&1&0&0&1&1&1&0 \\ 0&0&1&0&1&0&1&1 \\ 0&0&0&1&0&1&1&1 \end{array}\right] \quad (1)$$

$$P_{4\times4}{}^{T} = \left[\begin{array}{cccc} 1&1&1&0 \\ 1&1&0&1 \\ 0&1&1&1 \\ 1&0&1&1 \end{array}\right] \begin{array}{l} \leftarrow M_1 \\ \\ \\ \leftarrow M_2 \end{array} \quad (2)$$

$$G_2 = [\, I_{16\times16} \mid P_{16\times6} \,]$$

$$P_{16\times6}{}^{T} = \left[\begin{array}{cccccc} 1&1&1&0&1&1&1&0&1&1&1&0&1&1&1&0 \\ 1&1&0&1&1&1&0&1&1&1&0&1&1&1&0&1 \\ 0&1&1&1&0&1&1&1&0&1&1&1&0&1&1&1 \\ 0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1 \\ 0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1 \\ 1&0&1&1&0&1&0&0&0&1&0&0&1&0&1&1 \end{array}\right] = \left[\begin{array}{cccccccccccccccc} M_1&&&&M_1&&&&M_1&&&&M_1&&&\\ 0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1 \\ 0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1 \\ M_2&&&&\sim M_2&&&&\sim M_2&&&&M_2&&& \end{array}\right] \quad (3)$$

group a group b group c group d

check bits of an extended Hamming code H(22,16) with the generator matrix in Eq. (3), where $P_{16\times6}^{T}$ is the transpose of the parity matrix. The parity matrix of H(22,16) is constructed as follows: the first three rows in $P_{16\times6}^{T}$ of H(22,16) are duplications of the first three rows in $P_{4\times4}^{T}$ of H(8,4) (shown in Eq. (2)); the fourth and fifth rows in $P_{16\times6}^{T}$ are either four zeros or four ones; the last row in $P_{16\times6}^{T}$ is either the duplication of the last row in $P_{4\times4}^{T}$ or its inverse. The hardware implementation of the H(8,4) and H(22,16) encoder is shown in Fig. 3 and Fig. 4. A pattern of four ones in the fourth or fifth row requires the XOR of those four data bits, implemented by adding one extra XOR gate in each H(8,4) encoder, shown as the shaded XOR gate in Fig. 3.

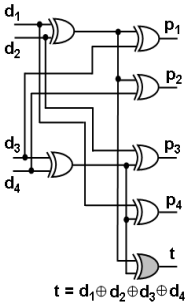

Fig. 3. H(8,4) encoder.

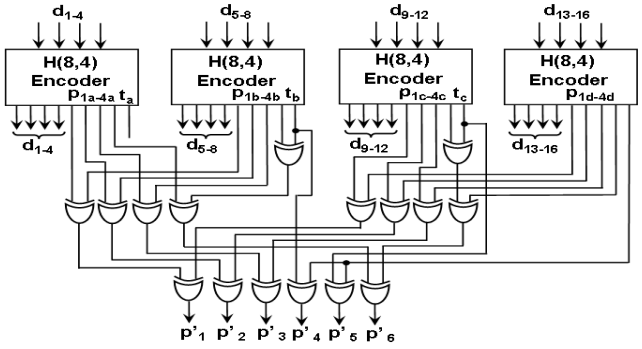

Fig. 4. Implementation of H(22,16) encoder.

3 Results and Analysis

In this section, the performance of the proposed dual-mode HARQ is evaluated. A 64-bit input message is used, which can be encoded using one extended Hamming code H(72,64) or four extended Hamming codes H(22,16). A 45 nm Predictive Technology Model (PTM) [7] is used and the link length is 3 mm with feature sizes from [8]. The clock frequency is 1 GHz. The residual flit error rate, which is the probability of error after decoding, is used to evaluate the reliability of the proposed method. The error probability of a single wire can be modeled by below [9],

$$\varepsilon = Q\left(\frac{V_{DD}}{2\sigma_N}\right) = \int_{\frac{V_{DD}}{2\sigma_N}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy \quad (4)$$

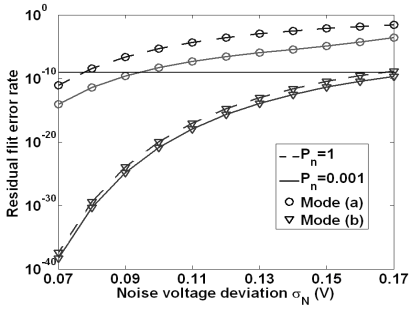

Fig. 5. Residual flit error rate.

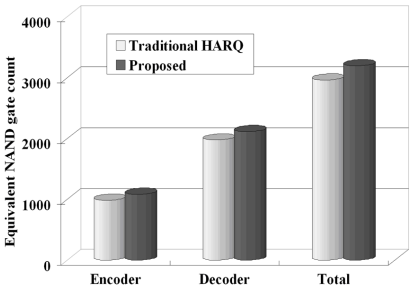

Fig. 6. Area comparison.

where V_{DD} is the supply voltage and σ_N is the standard deviation of the noise voltage, which is assumed to be a normal distribution. This model can be extended to describe multiple adjacent errors by assuming that a fault affects its neighboring wires with a certain probability P_n. Fig. 5 shows the residual flit error rate of the proposed dual-mode HARQ scheme working at different modes for different P_n values. The supply voltage is 1 V. The results show that mode (b) (separating into groups) achieves a significant improvement in residual flit error rate compared to mode (a) (traditional HARQ) when spatial burst errors are considered. For a 10^{-9} residual flit error rate requirement, mode (a) satisfies the requirement for σ_N<0.1. Mode (b) satisfied the requirement up to $\sigma_N \cong$0.17.

Fig. 6 compares the area of the proposed dual-mode scheme to a traditional HARQ with extended Hamming code H(72,64) in terms of equivalent NAND gate count. The results show that the codec area of the proposed method increases by about 10% compared to the traditional HARQ scheme. The codec power also increases about 10% compared to the traditional HARQ scheme. Fig. 7 shows the codec power and link power of the proposed method in each mode. The results show that link power dominates the total power consumption. Mode (b) consumes 18% more power compared to mode (a) because of the larger number of link wires used.

Fig. 8 evaluates the energy consumption of the proposed scheme for a residual flit error rate requirement of 10^{-9}. The results are compared to a traditional HARQ scheme using H(72,64) as well as worst-case noise design, in which four groups are always used. Two noise environments are considered. For the low noise environment (σ_N=0.07), the proposed dual-mode HARQ method works

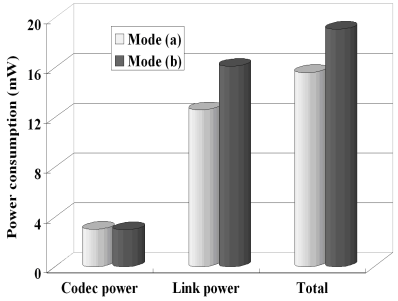

Fig. 7. Power consumption.

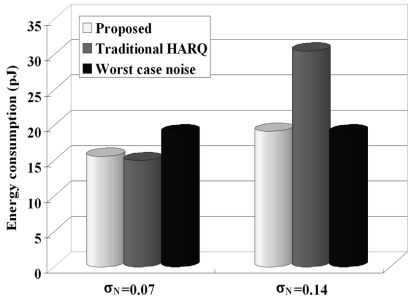

Fig. 8. Energy comparison.

in mode (a) and consumes similar energy to the traditional HARQ. Compared to worst- case noise design, the proposed method achieves about 18% reduction in energy consumption. In the high noise environment (σ_N=0.14), the proposed dual-mode HARQ scheme switches to mode (b). In order to meet this reliability requirement, traditional HARQ needs a higher link swing voltage [1, 10], which greatly increases the link energy. In the high noise environment, the proposed method consumes similar energy to worst-case noise design. Compared to previous solutions to increase the link swing voltage of the tradition HARQ scheme, the proposed dual-mode HARQ scheme can achieve 35% improvement in energy consumption.

4 Conclusion

In this paper, a dual-mode HARQ scheme is proposed for energy efficient on-chip communication. The efficient combination of a traditional HARQ scheme with interleaving shows a good balance between reliability and energy efficiency when burst errors are considered. In order to reduce codec area overhead, a hardware sharing design method is implemented and leads to a minor increase in area costs.

The type of error correction code can be dynamically selected in the proposed dual-mode HARQ scheme based on different noise environments. For a given system reliability requirement, the proposed error control scheme yields up to 35% energy improvement compared to previous solutions or up to 18% energy improvement compared to designing for worst-case noise.

References

1. Bertozzi, D., Benini, L.,De Micheli, G.: Error control schemes for on-chip communication links: the energy-reliability tradeoff. IEEE Trans. Computer-Aided Design Integr. Circuits Syst. 24, 818–831 (2005).
2. Murali, S., et al.: Analysis of error recovery schemes for networks-on-chips. IEEE Des. Test Comput. 22, 434–442 (2005).
3. Zimmer, H., Jantsch, A.: A fault model notation and error-control scheme for switch-to-switch buses in a network-on-chip. In: International Conference on Hardware/Software Codesign and System Synthesis, 188–193 (2003).
4. Sridhara, S., Shanbhag, N. R.: Coding for system-on-chip networks: a unified framework. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. 13, 655–667 (2005).
5. Li, L., Vijaykrishnan, N., Kandemir, M., Irwin, M. J.: Adaptive error protection for energy efficiency. In: IEEE/ACM International Conference on Computer-Aided Design, 2–7 (2003).
6. Rossi, D., Angelini, P., Metram, C.: Configurable error control scheme for NoC signal integrity. In: International On Line Testing Symposium, 43–48 (2007).
7. Arizona State Univ., Predictive Technology Model, http://www.eas.asu.edu / ptm/.
8. Xu, S., Benito, I., Burleson, W.: Thermal impacts on NoC interconnects. In: IEEE International Symposium on Networks-on-Chip, 220–220 (2007).
9. Hegde, R., Shanbhag, N. R.: Towards achieving energy-efficiency in presence of deep submicron noise. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. 8, 379–391 (2000).
10. Worm, F., et al.: A robust self-calibrating transmission scheme for on-chip networks. IEEE Trans. Very Large Scale Integr. (VLSI) Syst 13, 126–139 (2005).