Multi-Stage BCH Decoder to Mitigate Hotspot-Induced Bit Error Variation

Prashanthi Metku$^1$, Ramu Seva$^1$, Kyung Ki Kim$^2$ and Minsu Choi$^1$

$^1$Dept of ECE, Missouri University of Science & Technology, Rolla, MO, USA, \{pmcmc,rs2k6,choim\}@mst.edu
$^2$Dept of Electronic Eng., Daegu University, Gyeongsan, Korea, kkkim@daegu.ac.kr

Abstract—3D heterogeneous integration (commonly termed as 3DIC) of CPU, GPU and DRAM dies vertically interconnected by a massive number of TSVs (Through-Silicon Vias) is expected to overcome limited bandwidth, high latency and energy consumption of off-chip DRAM. However, spatial and temporal variability in temperature (i.e., hotspots) is anticipated to result in bit error variation in DRAM die. A novel multi-stage BCH decoder has been proposed to efficiently address this issue in this work. The proposed multi-stage BCH decoder is designed to tolerate up to a certain maximum number of error bits per codeword, which is estimated from the on-line thermal gradient data, to minimize the decoding latency.

I. INTRODUCTION

Processors are evolving toward a 3D heterogeneous integration (3DIC) of CPU, GPU and DRAM dies vertically interconnected by TSVs (Through-Silicon Vias) to alleviate power, bandwidth and latency bottlenecks. Fig. 1 shows an example where four heterogeneous dies (i.e., CPU, GPU, analog and DRAM) are stacked and interconnected by TSVs. When compared with the traditional off-chip interconnects, TSVs enable a massive number of vertical channels among CPU, GPU and DRAM dies while providing much shorter distance of data travel. Therefore, 3DHP (3D Heterogeneous Processor) technology is anticipated to inherently provide much higher bandwidth, low latency and power consumption. Despite numerous unprecedented benefits, however, there is a big challenge which is thermal reliability issues.

II. PRELIMINARIES AND REVIEW

Leakage power of DRAM cell is modeled to exponentially increase with temperature, $T$, as $P_{\text{leakage}} = P_0 \cdot \exp\left(\frac{-A}{T-B}\right)$ where $P_0$ is the room-temperature leakage power and parameters $A$ and $B$ are empirical constants [3], [4]. The number of discharged cells in DRAM is proportional to the dissipation of leakage power of cells. In [2], the relation between the error rate (i.e., the number of discharged cells divided by the total number of cells) and temperature is modeled as $E_{\text{DRAM}} \propto P_{\text{leakage}}$ where $E_{\text{DRAM}}$ is the error rate and $P_{\text{leakage}}$ is a function of $T$. The retention time distribution of cells is known to be divided into two regions: 1) tail distribution, and 2) main distribution [5]. For thermally-stable operation condition, the retention time of almost all the memory cells belong to "Main Distribution." However, there are a few memory cell whose retention time does not belong to "Main Distribution." This shorter retention time distribution is defined as "Tail Distribution." The refresh characteristics of DRAM are dominated by "Tail Distribution." Also, leakage power is exponentially increased as temperature rises, which means more and more

![Fig. 1: 3D stacking of CPU, GPU, analog and DRAM dies using TSVs [1].](image1)

![Fig. 2: DRAM error rate as a function of temperature and refresh period reported in [2].](image2)
cells fail to retain charge and become belonging to the "Tail distribution". To compensate this reduction in retention time, refresh period should be shorten to refresh more frequently. Yun et al have used the data reported in [6], [7] to fit the parameters and the resulting plot is shown in Fig. 2. This figure shows how DRAM error rate is related to temperature and refresh period.

III. THE PROPOSED MULTI-STAGE BCH DECODER

Hamming code [8] and Hsiao code [9] are widely used SECDED (Single-Error-Correction Double-Error-Detection) codes in memory products, which are efficient for correcting one error bit in a data word with acceptable area and performance overhead. As for multiple errors, codes with more parity bits are required. An example is the BCH (Bose-Chaudhuri-Hocquenghem) code, which is a class of powerful random error correcting cyclic codes [10] for multiple errors. Error correction (C) and detection (D) ability of an ECC code is a function of its Hamming distance \( d \) as \( d \geq C + D + 1 \) where \( D \geq C \). For example, SECDED code has \( d = 4 \) for \( C = 1 \) and \( D = 2 \). The Hamming distance of a code can be increased by providing more parity bits. The code rate is defined as the fraction \( \frac{k}{n} \) of \( k \) data bits, \( n \) encoded bits and \( r = n - k \) parity bits. For SECDED Hsiao code, the length of parity can be determined by \( 2^r + 1 \geq k + r \).

One particularly famous SECDED code for DDR DRAM application is \((72,64)\) code where \( k = 64 \) and \( n = 72 \), since nine 8-bit wide DRAM chips can be configured to offer it without any wasted bits.

![Fig. 3: The proposed multi-stage BCH decoder.](image)

The proposed BCH decoder consists of four decoding path as shown in Fig. 3. Each is designed to detect and correct up to a specific maximum number of error bits. The temporal and special changes in the temperatures are recorded by the temperature sensors, from which the BEPs are calculated. Assuming that there will be maximum number that many errors for that particular range of BEP values, the respective BCH decoder is selected. Therefore BEP is the main decision maker block in the dynamic BCH decoder. No decoding operation is needed when the calculated BEP is zero, since the received codeword is expected to contain no error bits. This no-decoding path is represented as BCH0 in Fig. 3. When the expected maximum BEP is 1, BCH1 block is selected which can detect and correct up to one error bit. Similarly, BCH2 block can be selected to detect and correct up to 3 errors, BCH3 up to five error bits and BCH4 up to seven error bits. Each decoder has different area overhead and decoding latency proportional to its error correction power. From the on-line thermal gradient data, BEP is estimated. By varying BEP, the probability of number of error occurred can be calculated using Binomial Distribution as \( P(k) = \frac{n!}{k!(n-k)!} \cdot p^k \cdot (1-p)^{n-k} \), where \( n = 511 \) (i.e., word size of the proposed decoder), \( k = \# \) of error bits, \( p = \) Bit Error Probability (BEP). From this equation, it is possible to calculate the distribution of the expected number of error bits per codeword.

A BCH decoder consists of three decoding steps. They are Syndrome Calculator, Berlekamp-Massey (BM) algorithm and Chien Search [11]. These steps are inter-related as the syndrome calculator calculates the syndrome according to the received data, from these values error locator polynomial is generated, in turn error location is calculated from these polynomial by the Chien search method.

IV. SIMULATION RESULTS

The advantage of multi-stage BCH decoder to single stage BCH decoder is its ability to adapt to BEP variance for reduced decoding latency/power. An FPGA-based prototype decoder has been designed and its resource utilization and decoding latency results are summarized in Table I. The results shown in this table clearly shows the expected trade-off between the maximum number of corrected error bits per codeword and area overhead & decoding latency. For example, if the estimated maximum error bits per codeword is 5, then BCH3 path should be selected to correct error bits up to 5 and its decoding process takes 28.59ns.

<table>
<thead>
<tr>
<th>BCH1</th>
<th>BCH2</th>
<th>BCH3</th>
<th>BCH4</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Slice Regs</td>
<td>49</td>
<td>93</td>
<td>220</td>
</tr>
<tr>
<td># of Slice LUTs</td>
<td>107</td>
<td>586</td>
<td>1817</td>
</tr>
<tr>
<td># of LUT FFs</td>
<td>44</td>
<td>98</td>
<td>203</td>
</tr>
<tr>
<td>Decoding latency</td>
<td>7.21ns</td>
<td>12.73ns</td>
<td>28.59ns</td>
</tr>
</tbody>
</table>

REFERENCES


