DART: Dynamic Repair for Interconnect Fault Tolerance in Hybrid Bonding*

Partho Bhoumik †, Zhichao Chen‡, Puneet Gupta‡ and Krishnendu Chakrabarty†
†School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ
‡Department of Electrical and Computer Engineering, University of California, Los Angeles

***Abstract*—Chiplet-based heterogeneous integration has become a cornerstone of next-generation semiconductor packaging, offering high-bandwidth and energy-efficient solutions. Hybrid bonding for heterogeneous integration enables ultra-dense chiplet-to-chiplet interconnects but introduces significant reliability challenges. Manufacturing defects and in-field degradation can generate clustered or line-shaped fault patterns that cannot be effectively addressed by conventional repair methods. We introduce a cluster-aware repair methodology, referred to as dynamic repair for interconnect fault tolerance in hybrid bonding (DART), that leverages non-uniform, irregular repair chains to tolerate both clustered and line defects with minimum rerouting overhead. A multi-objective heuristic optimization framework is developed to construct these irregular chains efficiently while satisfying performance constraints. Comprehensive design-space exploration across diverse bump grids, chain structures, and spare ratios demonstrates the robustness of the proposed approach. Results show that DART improves repairability up to 50% over state-of-the-art methods for clustered defects, with comparable or lower overhead.**

I. Introduction

The increasing demand for compute and memory-intensive workloads in artificial intelligence and high-performance computing is straining traditional monolithic system-on-chip designs. However, further scaling of die size is constrained by the fixed reticle field limit of lithography systems. As die sizes approach or exceed this limit, the need for die stitching increases, along with the probability of yield-limiting defects. To overcome these constraints, the industry is shifting toward chiplet-based integration, where multiple smaller dies are fabricated independently and interconnected at the package level. This paradigm is enabled by advanced packaging technologies, which support high-density, fine-pitch interconnects between heterogeneous dies [1]–[3].

Among these technologies, hybrid bonding (HB) has emerged as a leading approach for achieving extremely high I/O density through direct dielectric-to-dielectric and metal-to-metal bonding between die surfaces, often at sub-10 μm pitch [4], [5]. While this method significantly enhances system performance, bandwidth, and modularity, the ultra-fine pitch and compact interconnect geometry introduce new reliability challenges. The dense arrangement of I/O bumps makes hybrid-bonded interfaces more susceptible to manufacturing-induced defects such as voids, misalignments, and particle-induced opens or shorts [6], [7]. These failures frequently exhibit clustered or line-shaped defect patterns [5], [8], especially in regions of dense routing or non-uniform stress. Such spatially correlated defects present major obstacles for conventional interconnect repair architectures, which typically assume random, independent defect distributions.

The repair mechanism defined in the Universal Chiplet Interconnect Express for advanced packages (UCIe-A) and the Advanced Interface Bus (AIB) standards primarily target sparsely distributed defects [9], [10]. Repair frameworks for clustered defects have been proposed for through-silicon via (TSV)-based 3D integration [11]. They are effective for relatively coarse-pitch interconnects, but they are not designed to scale to the sub-micron regimes typical of HB. At these finer pitches, the physical proximity of bumps and the increased likelihood of spatially correlated failures demand repair strategies that are both cluster-aware and topology-adaptive.

Moreover, while clustered defects are often introduced during manufacturing, small voids or weak interconnects that escape initial test can degrade over time due to electromigration and thermomechanical stress [12]–[14]. Such progressive degradation may eventually trigger in-field failures, emphasizing the need for on-chip spare interconnect provisioning and dynamic, in-field repair mechanisms capable of reconfiguring signal paths around degraded interconnects. Therefore, developing lightweight yet robust repair architectures tailored to the spatial and temporal failure modes of ultra-dense interconnect fabrics remains an open and critical research direction.

To address these reliability challenges, we propose dynamic repair for interconnect fault tolerance in hybrid bonding (DART), which forms multiple irregular repair chains to reroute faulty bumps caused by both clustered and line-shaped defect patterns. DART employs heuristic algorithms to construct repair chains under a given spare budget and to intelligently place spare bumps so that defective interconnects within a cluster can be efficiently bypassed by rerouting to available spares. We further present a lightweight algorithm that dynamically maps faulty signals to spare interconnects on-chip without incurring any additional PHY-layer area overhead beyond what is already provisioned in existing UCIe or AIB. We show how designers can use DART to select appropriate repair-chain configurations and spare ratios to effectively target specific cluster sizes and defect distributions. The main

*This research was supported by CHIMES, one of the seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

979-8-3315-6337-0/26/$31.00 ©2026 IEEE

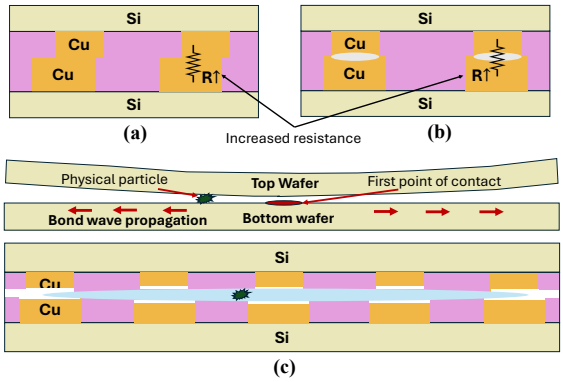

Fig. 1: Cause of defects during HB: (a) alignment; (b) Cu recess; (c) particle-induced [5].

contributions of this work are summarized as follows:

- We introduce the first cluster-aware repair framework tailored for hybrid-bonded packages, capable of addressing both localized clusters and line-shaped defect patterns.
- We propose a graph-coloring and simulated-annealing (SA)-based heuristic to generate zig-zag repair chains that maximize rerouting flexibility.
- We develop multi-objective optimization strategies that jointly improve repairability and minimize performance degradation due to extended signal paths.
- We present a hardware-efficient on-chip repair mechanism that dynamically remaps faulty bumps to available spares.

II. Background

A. Defects in HB

HB facilitates direct Cu-Cu bonding at ultra-fine pitches. Unlike fanout and interposer-based packages, HB does not require any redistribution layers to route signals, allowing for much tighter bump pitches and a significantly higher bump density. However, as pitch scales down, the margin for alignment and planarity shrinks, requiring sub-micrometer accuracy in wafer alignment, surface preparation, and bonding. Even with advanced bonding tools and stringent cleanroom protocols, defects can compromise interconnect yield and reliability. Some of them are detailed below.

Misalignment: Misalignment occurs when opposing Cu pads on two dies do not align during bonding; see Figure 1(a). This defect can arise from tool calibration errors, die shift or rotation, and wafer warpage induced by coefficient of thermal expansion mismatch or chemical–mechanical polishing (CMP) non-uniformity [5], [7], [15], [16]. With increasing offset, misalignment reduces Cu–Cu contact, causing resistive opens or complete disconnection between dies. Localized warpage or Cu recess dishing can cause local misalignment, creating clusters of faulty bumps while other areas remain functional. In contrast, when the error stems from global alignment issues such as robotic arm drift, the entire die interface may be affected, making repair infeasible because spares may be compromised too.

Voids: One major cause of void formation is excessive Cu recess following CMP, which leaves concave or uneven copper pad topography, as shown in Figure 1(b). In addition, insufficient Cu recess or protruding copper can generate localized stress peaks at the interface during anneal, leading to delamination or cracking of the dielectric layer. Surface roughness and local density variations further aggravate this by altering planarity and leading to incomplete bonding and interfacial voids [5]. Physical contaminants from wafer dicing, CMP residues, or ambient particles can act as obstacles between the die surfaces, forcing the bond wave to circumvent or split around them; see Figure 1(c). The result is a "main void" trailing a "void tail" behind the particle, extending radially across the interface [5], [8]. Because these voids extend along the bond front, they cause clustered or line-shaped fault patterns rather than isolated single-pad failures.

Shorts: As reported by [17], excessive Cu diffusion across the bonding interface can lead to the formation of metallic bridges between neighboring pads, resulting in electrical shorts. [18], [19] showed that residual conductive contaminants or incomplete surface cleaning can promote Cu–Cu bridging after post-bond annealing, particularly at sub-micrometer pitches.

In summary, failures in HB rarely occur in isolation, but rather emerge as spatially correlated defect regions, making localized, topology-aware repair essential.

B. Related Prior Work

Chiplet-to-chiplet interconnect standards have proposed repair schemes based on rerouting signals along predefined chains. When a bump is found to be faulty in these chains, its signal is sequentially shifted to adjacent bumps until a spare is reached. AIB introduces two spare bumps placed near the center of a repair chain. Each signal S_n can be routed to its nominal bump B_n or to neighboring bumps, either B_{n-2} or B_{n+2}, depending on its position relative to the spares. During repair, signals are shifted left or right until one of the central spares is utilized [10]; see Figure 2(a).

UCIe-A adopts a similar approach but places spares (one spare per 16 data bumps) at the two ends of each chain [9]. Signals can be rerouted in either direction, using the left spare for the first defect and the right spare for a second. However, because these repair chains are regular and linear, even a small 2×2 clustered defect cannot be repaired under this scheme. UCIe-3D addresses clustered defects by partitioning the bump grid into modules and reserving several as spares [20]. When a fault is detected in an active module (e.g., d_0), the entire module is rerouted to a corresponding spare (s_0); see Figure 2(b). This approach incurs large hardware overhead (s_0 needs multiplexing between $d_0, d_3, m_0, m_2, m_4, d_{13}, d_{14}$, requiring up to a 7-to-1 fan-in in the PHY), which introduces routing complications and delay. The method is also fragile, as repairing d_0 leaves in-field faults in other modules (e.g., d_3) unrepairable, and any defect inside a spare module eliminates the repair capacity for all modules mapped to it. As a result, UCIe-3D is not robust to random failures, nor is it effective for void-tail fault lines that cut across multiple module boundaries.

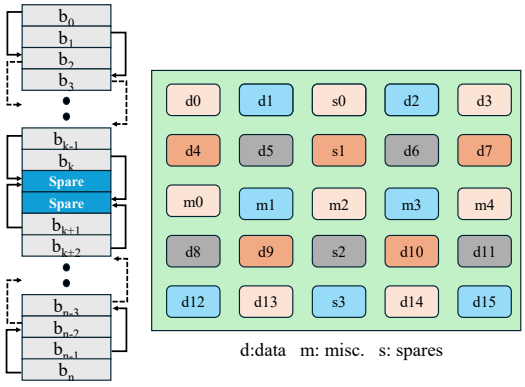

Fig. 2: Repair techniques for (a) AIB; (b) UCIe-3D.

UCIe-A derivative techniques have also been explored to improve baseline repair rates focusing on shorts [21]–[23]. In [21], the authors construct repair chains by clustering bumps that cannot short to one another and allocating a single spare bump to each chain. They apply graph coloring to ensure neighboring bumps are not grouped together, and report that each chain can reliably handle only about 4-5 defects. However, this approach is not well suited for aggressively scaled HB packages as open defects dominate here and combining physically distant bumps into the same chain can introduce substantial routing delay. Clustered-defect repair schemes have also been proposed for TSV-based 3D integration, e.g., ring, router, group, cellular, and honeycomb architectures [24]–[28]. However, these approaches target sparse TSV arrays where large defect clusters are unlikely, and they require substantially higher spare ratios (often one spare per 10 signals or less).

III. Proposed Methodology

Our goal is to develop a repair strategy that tolerates clustered and line-shaped defect patterns while retaining the ability to repair random, isolated defects. To meet both objectives, we avoid relying on spare-module or spare-cluster mechanisms and instead build upon the conventional repair-chain paradigm. However, a fundamental limitation of existing approaches is that traditional repair chains are regular and fully localized such that each chain occupies an entire contiguous region of the bump map. This means that once a large defect cluster appears inside that region, only a single repair chain is responsible for rerouting all the faulty bumps. Such a concentration of defects can easily exceed the chain's available spare capacity, and in many cases the cluster may also damage the chain's own spare bumps, further diminishing repairability.

To overcome this limitation, we propose irregular, zig-zag repair chains that are intentionally spatially interleaved rather than confined to isolated regions. By distributing each chain across different areas of the bump grid, any region that experiences a defect cluster will intersect with multiple repair chains instead of just one. As a result, defects within the cluster become naturally partitioned among several chains, allowing each chain to reroute only a fraction of the defects and significantly increasing the probability of successful repair.

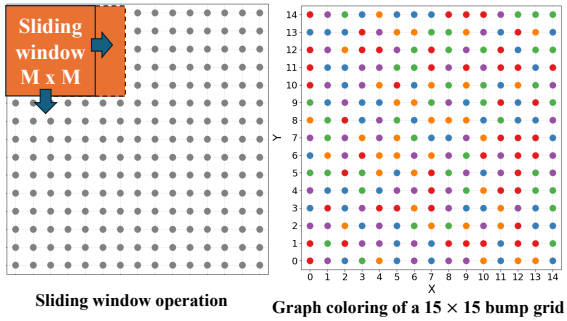

Fig. 3: Bump-to-chain assignment using graph coloring.

A. Repair Chain Formation

Achieving spatial dispersion requires assigning adjacent bumps within a region to different repair chains. We formulate the bump-to-chain assignment as a graph-coloring problem, where each color represents a repair chain. By enforcing high color diversity within each local neighborhood, we ensure that each region of the bump grid includes multiple chains while respecting the total number of chains permitted by the spare-budget constraints. We define the bump map as an $N \times N$ grid $\mathcal{G} = \{g_{i,j}\}$, where each bump $g_{i,j}$ is assigned one of K colors, i.e., $g_{i,j} \in \{0, 1, \ldots, K-1\}$.

Our goal is to assign these K colors such that every $M \times M$ window contains as many distinct colors as possible. Here, K is the number of available repair chains, and M is a design parameter chosen for the target cluster size. To enforce this requirement, we use an $M \times M$ sliding window over the grid with stride 1; see Figure 3. Initially, each window randomly assigns colors to its bumps. As the window slides, any unassigned bump selects a color not already used within that window. This policy maximizes local color diversity and produces spatially interleaved, irregular repair chains. However, maximizing diversity alone can force bumps of the same color to appear far apart, leading to fragmented chains. Such chains require long detours during rerouting, incurring delay and harming signal integrity, an undesirable property for high-speed chiplet interconnects. Thus, in addition to promoting diversity, each color class must remain spatially compact, forming a chain whose bumps are close together. To jointly enforce these properties, we define two objective functions:

A) Global Diversity Penalty

For each sliding window $W_{pq} \subset \mathcal{G}$ of size $M \times M$ with top-left corner at (p, q):

$$W_{pq} = \{\, g_{i,j} \mid p \le i < p+M,\ q \le j < q+M \,\}.$$

Let U_{pq} be the number of unique colors in that window. The diversity penalty for window (p, q) is: $l_{pq}^{\text{div}} = (M^2 - U_{pq})$. The total diversity loss over the grid (stride = 1) is:

$$\mathcal{L}_{\text{div}} = \sum_{p=1}^{N-M+1} \sum_{q=1}^{N-M+1} l_{pq}^{\text{div}}$$

A small value of $\mathcal{L}_{\text{div}}$ indicates that each local region contains many distinct colors, ensuring that multiple repair chains pass through any potential defect cluster.

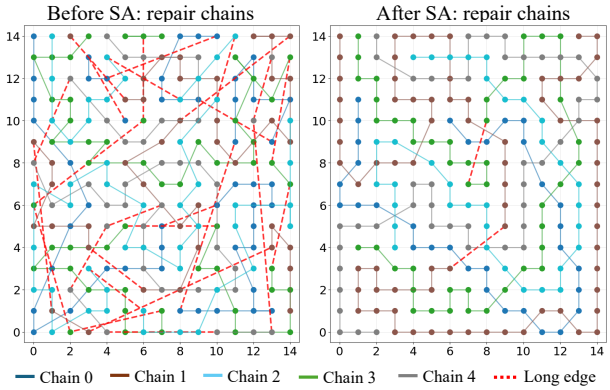

Fig. 4: Repair chain formation before and after SA optimization.

B) Spatial Fragmentation Penalty

For each chain (color) $c \in \{0, 1, \ldots, K-1\}$, we define the set of bump positions: $P_c = \{p_1, p_2, \ldots, p_{n_c}\}$, where $p_i = (x_i, y_i)$. We compute pairwise Euclidean distances:

$$d_{ab}^{(c)} = \|p_a - p_b\|_2 = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

To approximate the spatial connectivity of each chain, we use a greedy nearest-neighbor traversal. First, we choose a random starting bump $p_t \in P_c$. At each step, move to the nearest unvisited bump: $p_{t+1} = \arg\min_{q \in Q_t} \|q - p_t\|_2$, where Q_t is the set of unvisited bumps. The total traversal length for a particular chain (of color c) is, $l_c = \sum_{t=0}^{n_c-2} \|p_{t+1} - p_t\|_2$. The overall fragmentation (compactness) penalty is $\mathcal{L}_{\text{frag}} = \sum_{c=0}^{K-1} l_c$. Lower $\mathcal{L}_{\text{frag}}$ indicates that bumps of the same color are spatially close, forming compact chains that minimize rerouting delay. By minimizing these two metrics, we can balance defect tolerance and reroute delay, producing repair chains well-suited for hybrid-bonded interconnect fabrics.

B. Repair Chain Optimization

Our overall objective is to jointly minimize the diversity and fragmentation losses, $\mathcal{L}_{\text{div}}$ and $\mathcal{L}_{\text{frag}}$. Although this problem can be formulated as integer linear programming (ILP) to find an exact solution, the resulting model becomes unwieldy for realistic bump-map sizes. The color assignment itself introduces N^2K binary variables $x_{i,j,c}$. Enforcing color diversity across all sliding $M \times M$ windows adds another $(N-M+1)^2K$ variables $y_{p,q,c}$ to indicate whether each color appears in each window. The primary source of complexity comes from modeling chain compactness: capturing adjacency or ordering within each color class requires defining pairwise variables z_{c,p_1,p_2} over all N^2 bump positions, resulting in KN^4 binary variables. This leads to millions of variables even for modest grids (e.g., $N = 20$). Such formulations exceed the limits of practical ILP solvers. Therefore, we adopt a heuristic SA approach, which can efficiently optimize both $\mathcal{L}_{\text{div}}$ and $\mathcal{L}_{\text{frag}}$ without incurring the complexity of ILP.

Our objective is to minimize the combined loss $E(\mathcal{G}) = w_{\text{div}}\,\mathcal{L}_{\text{div}} + w_{\text{frag}}\,\mathcal{L}_{\text{frag}}$, where w_{div} and w_{frag} control the relative importance of local color diversity and chain compactness. At each SA iteration, a new configuration $\mathcal{G}'$ is generated by randomly swapping the colors of two bumps. The corresponding change in energy is computed as $\Delta E = E(\mathcal{G}') - E(\mathcal{G})$. If $\Delta E < 0$, the new configuration is accepted; otherwise, it is accepted with probability $\exp(-\Delta E/T)$, where T denotes the current annealing temperature. This probabilistic acceptance policy enables the search to escape local minima and progressively converge toward a low-energy solution.

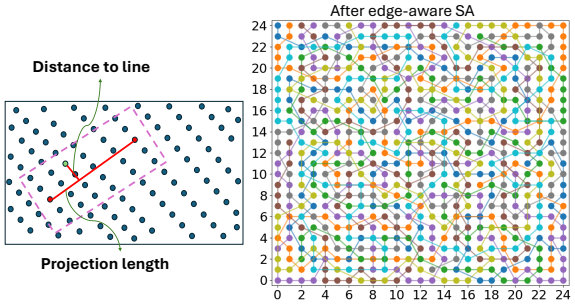

Fig. 5: Repair chains designed (for N= 25) using edge-aware SA.

The initial configuration $\mathcal{G}_0$ is generated using the sliding-window greedy method described earlier. SA then iteratively refines this initialization to obtain the final configuration $\mathcal{G}^* = \arg\min_{\mathcal{G}} E(\mathcal{G})$. Figure 4 demonstrates how SA improves the compactness of repair chains in a bump grid. Without SA, each chain is fragmented into several disconnected pieces scattered across the layout. SA significantly reduces this fragmentation, producing tighter and more contiguous chains. However, some outlier edges still remain (highlighted in red), where two consecutive bumps are placed far apart. Such long edges are physically unrealistic and undesirable. To address this limitation, we introduce Edge-Aware Simulated Annealing, a modified SA formulation in which recoloring moves are no longer random but strategically guided toward eliminating long edges that arise after graph coloring.

C. Edge-Aware Simulated Annealing

We incorporate an additional constraint that guides the color reassignment toward reducing long edges in the greedy chain traversal. Let $\mathcal{E}_c$ denote the set of consecutive edges in the greedy path for chain c. We define the set of long edges as

$$\mathcal{E}_{\text{long}} = \bigcup_{c=0}^{K-1} \{(p_a, p_b) \in \mathcal{E}_c \mid \|p_a - p_b\|_2 > \tau\},$$

where τ is a user-defined length threshold. Each long edge indicates a region where the chain must be locally "repaired" by relocating or recoloring bumps to create intermediate connections. To identify candidate bumps for recoloring, we construct an oriented bounding box around the line segment joining the endpoints of a long edge. For every bump inside this region, we compute the projection length of the bump onto the line segment, and the perpendicular distance from the bump to the line, as shown in Figure 5.

A bump is considered a valid recoloring candidate if (i) its projected position onto the edge lies within the segment, $0 < \text{proj\_length} < L$, (ii) it is sufficiently close to the edge, $\text{dist\_to\_line} < d_{\max}$, and (iii) its current color differs from the target color. Here, L is the edge length and $d_{\max}$ is a

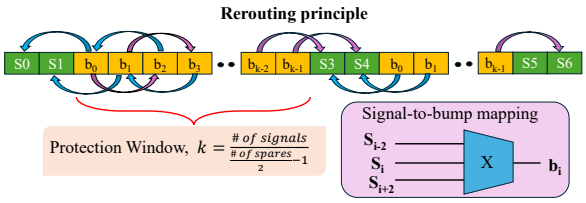

Fig. 6: DART rerouting scheme. Signals can shift in either direction to access the paired spares placed after every k bumps in the chain.

small proximity threshold. These constraints ensure that only bumps positioned appropriately along and near the long edge are selected. During SA, recoloring moves are guided toward these candidates rather than chosen arbitrarily. Figure 5 shows the resultant repair chains formed by edge-aware SA.

D. Repair Chains and Spare Allocation

The task now is to reroute the repair chains efficiently. Consider a chain containing n consecutive signal bumps $\{b_0, b_1, \ldots, b_{n-1}\}$. In traditional UCIe repair, all n signals share 2 spares placed at the ends of the chain, which leaves the structure highly vulnerable to clustered defects. To address this limitation, DART utilizes m paired-spare blocks in each chain (of n bumps), where each block provides two spares. The total number of spares ($2m$) is determined by the redundancy budget, and the paired-spare blocks are inserted at the beginning of the chain and after every $k = \lceil n/(m-1) \rceil$ signal bumps; see Figure 6. The motivation for using paired spares instead of single spares arises from how effective a value of k is for clustered failures. In a single-spare layout, each spare covers only its nearest k signals, and as more spare locations are inserted, k becomes smaller. This creates many short protection windows, making the spare bumps themselves vulnerable to the same defect clusters affecting nearby signals. Pairing the spares, however, effectively halves the number of protection windows while keeping the same total spare count. As a result, each window becomes twice as large, the spares are spread farther apart, and every window gains access to four spares (two from the left paired block and two from the right). Thus, without increasing design cost, the paired-spare configuration provides significantly higher tolerance to spatially correlated defects.

We adopt a second-adjacent reroute rule inspired by AIB; each bump b_i receives inputs $\{S_i, S_{i-2}, S_{i+2}\}$. When a defect is localized, the affected signal is rerouted to the next valid bump while skipping the immediate neighbor. Thus, the MUX requirement in the PHY layer stays up to 3-to-1 MUX, which is similar to that of UCIe-A. Under this rule, all odd-indexed (even-indexed) bumps map consistently to one of the two spares on the left (right) and right (left), creating a balanced and conflict-free routing structure.

E. Repair Algorithm

Each bump in a repair chain is indexed, and thus an interconnect repair language can be used to preset the logic during the design phase [23]. When a defective bump b_i is detected, the algorithm first checks whether the left paired-spare block is available. If so, it initiates a left-shift operation

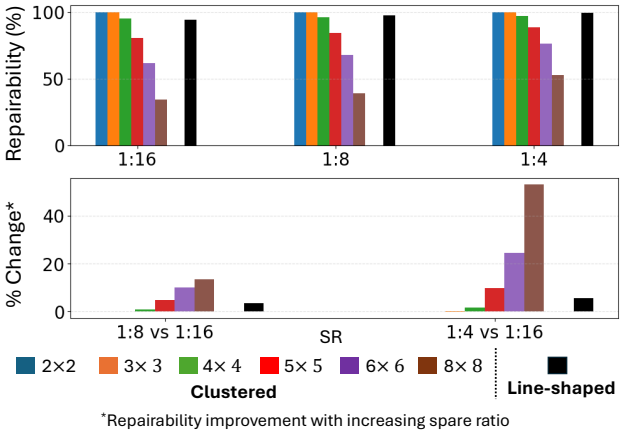

Fig. 7: Repairability analysis on a 25×25 bump grid with 8 repair chains and multiple SRs in case of clustered and line-shaped defects.

by programming the MUXes of a sequence of bumps b_j, where $j = i - 2p$. For each such bump, the MUX is configured to select the signal arriving from b_{j+2}, effectively propagating the signal leftward through the chain until it reaches the left spare. If the left spare is unavailable or if another defect appears deeper within the same chain, the algorithm instead initiates a right-shift operation. In this mode, the select bumps follow $j = i+2p$, and each MUX selects the signal from b_{j-2}, driving it toward the right-paired spare. Since each fault is mapped to its chain in constant time and each repair updates at most $k/2$ interleaved positions within the protection window, the overall time cost for repair per fault is linear in k, i.e., $O(k)$.

Overall, DART integrates four complementary optimization techniques to maximize resilience against clustered defects. First, the irregular chains-generation process ensures that more repair chains are routed through localized defect-prone regions. Second, the edge-aware SA enhances chain diversity while constraining reroute delay. Third, intelligent spare allocation places spares in paired blocks, providing higher redundancy without increasing design cost. Finally, the second-adjacent rerouting rule implicitly partitions each chain into two interleaved subchains (odd and even-indexed signals), each mapping to its own pair of spares, thereby increasing the number of effective repair chains within a local space.

IV. Experimental Results

We evaluate the effectiveness of DART under both clustered and line-shaped defect patterns. A defect cluster is first placed at the top-left corner of the bump map. Next it is moved horizontally and vertically with a stride of one. This allows every possible placement of the defect to be analyzed. Repairability is computed as the ratio between the total number of repaired faulty bumps and the total number of signal faults across all cluster positions. For line-shaped defects, we begin by drawing a circle centered at the middle of the bump map. Lines are then drawn within this circle, each separated by one degree. The bumps that lie near these radial lines are marked as faulty for each scenario, and the repair rate is computed accordingly.

Figure 7 quantifies repairability for different cluster sizes on a 25 × 25 bump map with eight repair chains and M = 3. With

a spare ratio (SR) of 1:16, DART achieves 100% repairability for a 2 × 2 cluster. Even for a 5 × 5 cluster, the repairability remains high at 94.45%. As the cluster becomes larger, the performance begins to decline. For an 8 × 8 cluster, the repair rate drops to 56%. This decline occurs because the number of available spares is insufficient to handle such large clusters. Increasing the SR improves repairability for large clusters. With a ratio of 1:4, repairability increases to 69% for an 8 × 8 cluster. However, large clusters also damage many of the spare bumps, which limits the improvement. In fact, for some bump maps, increasing the SR can slightly reduce repairability. For small clusters up to 4 × 4, more spares are compromised by the defect. Figure 7 also shows the repairability for line-shaped defects using the same bump map and repair-chain settings. With an SR of 1:16, the repairability is 94.13%, indicating that most signal paths can still be redirected to a valid spare. When the SR increases to 1:4, the repairability reaches near 100%. The paired-spare layout effectively handles both compact defect clusters and extended line-shaped defects, provided that sufficient spare capacity is available.

Next, we evaluate how the sliding-window size affects repairability for different bump-grid dimensions. We consider grids with N = 15, 20, 25, and for each grid, we evaluate sliding-window sizes $M = 3$ and $M = 5$. Clustered defects are modeled in terms of a 5×5 cluster, while line-shaped defects are generated using 1° radial separations. The results are summarized in Table I. We note that repairability is generally higher when the sliding-window size is smaller. A small window allows greater diversity during chain formation, whereas a larger window reduces the effective diversity as K is fixed and $K \ll M^2$, ultimately degrading defect tolerance.

There is one notable exception: the 20-bump grid with $M = 3$, where performance drops relative to $M = 5$. This occurs due to the specific geometry of the repair chains, in which several radial fault lines intersect regions with paired spares more aggressively, causing a larger fraction of spares to be faulty. Although our design flow initially selected M based on the expected cluster size, these results suggest that smaller sliding-window sizes tend to improve overall diversity and yield more fault-tolerant designs. We also observe that larger bump grids achieve higher repair rates. This trend is encouraging, as future chiplets are expected to scale in I/O count, naturally resulting in larger bump maps. For extremely large grids, the bump array can be partitioned into smaller local regions; repair chains can then be generated for a representative region and replicated across the remaining regions.

We compare the repairability of DART with state-of-the-art redundancy schemes designed for clustered defects. [28] conducted an extensive comparison with previously proposed repair methods for TSV-based 3D architectures [24]–[27]. The authors consider a cluster window of size z×z and assume that an arbitrary subset of TSVs inside that window can become defective. They report results for up to nine defects within the windows for z = 6, 5, 4, and 3. Since our model assumes that every bump within a cluster can fail simultaneously, the closest comparable scenario from [28] is the case z = 4 with nine faults. [21] presented the results for a maximum of six defects due to three realistic shorts between neighboring bumps. We compare these results with DART, where 3×3 clustered defects are injected into a 15×15 bump grid with SR of 1:16. The results are summarized in Table II.

TABLE I: Impact of sliding window size on repairability.

Design (N, K, M)	**Repair rate for clustered defects (%)**	**Repair rate for line defects (%)**
15, 5, 3	80.62	91.30
15, 5, 5	69.93	91.20
20, 7, 3	91.32	92.4
20, 7, 5	86.35	92.91
25, 8, 3	94.45	94.13
25, 8, 5	90.13	93.59

TABLE II: Comparison between repair schemes in terms of repairability and cost.

Scheme	**SR**	**# Faults**	**Repair rate(%)**	**Area (μm^2)**
Ring [24]	1:7.25	9*	10	335.6
Router [25]	1:5.24	9*	90	1036.8
Group [26]	1:7.25	9*	50	376
Cellular [27]	1:10	9*	8	388.6
Honeycomb [28]	1:6.39	9*	53	496
Chuang et al., [21]	1:16	6**	48.6	163.2
DART	1:16	9 (3×3)	**99.71**	367.2

*Nine faults are considered in a 4×4 cluster window.
**All possible cases of 3 shorts in a physical-aware bump map.

DART achieves a substantial increase in repairability for all cases. Whereas prior TSV-based schemes repair only 50–53% of nine clustered defects (cellular/ring-based schemes drop to 8–10%), DART consistently repairs more than 99% of all 9-fault cluster scenarios. Even the router-based design, which posts the strongest performance among previous methods, repairs only 90% of defects while consuming over 3× more area and requiring nearly double SR. We also compare area overhead using the same cell-library area models reported in [28]. For DART and [21], we compute the area under a common configuration of 64 signal bumps, and four spares to match the normalized analysis in [28]. DART's rerouting logic introduces only a modest area cost, comparable to the honeycomb architecture [28] and noticeably lower than the router-based scheme. [21] exhibits the smallest area because its repair logic is unidirectional (either left or right) and hence uses only 2-to-1 MUXes, but this also results in substantially lower repairability. In contrast, DART provides near-perfect repair coverage while maintaining area efficiency close to low-overhead designs and significantly better than high-overhead solutions.

V. Conclusion

We have presented a cluster-aware repair framework (DART) to tolerate both clustered and line-shaped defects in hybrid-bonded interfaces. By incorporating simulated annealing, DART improves local chain diversity and enhances bump-level consistency within each chain. The resulting lightweight and intelligent rerouting mechanism achieves significantly higher repairability than state-of-the-art methods while requiring minimal spare resources.

REFERENCES

[1] S. Li *et al.*, "High-bandwidth chiplet interconnects for advanced packaging technologies in AI/ML applications: Challenges and solutions," *IEEE Open Journal of the Solid-State Circuits Society*, 2024.

[2] S. Chen *et al.*, "The survey of 2.5D integrated architecture: An EDA perspective," in *Proceedings of the 30th Asia and South Pacific Design Automation Conference*, 2025, pp. 285–293.

[3] J. H. Lau, "Recent advances and trends in advanced packaging," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 12, no. 2, pp. 228–252, 2022.

[4] C.-K. Hsiung and K.-N. Chen, "A review on hybrid bonding interconnection and its characterization," *IEEE Nanotechnology Magazine*, vol. 18, no. 2, pp. 41–50, 2024.

[5] Z. Chen and P. Gupta, "YAP: Yield modeling and simulation for advanced packaging," in *62nd ACM/IEEE Design Automation Conference (DAC)*, 2025, pp. 1–7.

[6] V. H. Vartanian *et al.*, "Metrology needs for 2.5D/3D interconnects," *Handbook of 3D Integration*, pp. 393–430, 2014.

[7] T. Workman *et al.*, "Die to wafer hybrid bonding and fine pitch considerations," in *IEEE 71st electronic components and technology conference (ECTC)*, 2021, pp. 2071–2077.

[8] F. Nagano *et al.*, "Void formation mechanism related to particles during wafer-to-wafer direct bonding," *ECS Journal of Solid State Science and Technology*, vol. 11, no. 6, p. 063012, 2022.

[9] D. D. Sharma *et al.*, "Universal chiplet interconnect express (UCIe): An open industry standard for innovations with chiplets at package level," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 12, no. 9, pp. 1423–1431, 2022.

[10] D. Kehlet *et al.*, "Accelerating innovation through a standard chiplet interface: The advanced interface bus (AIB)," *Intel White Paper*, 2017.

[11] S.-C. Hung *et al.*, "Design-for-test solutions for 3-D integrated circuits," *Integrated Circuits and Systems*, vol. 1, no. 1, pp. 3–17, 2024.

[12] P. Bhoumik *et al.*, "Fault modeling and testing of chiplet-to-chiplet interconnects in fan-out wafer-level packaging*," in *IEEE International Test Conference (ITC)*, 2025, pp. 357–366.

[13] S. Moreau *et al.*, "Hybrid bonding-based interconnects: A status on the last robustness and reliability achievements," *ECS Journal of Solid State Science and Technology*, vol. 11, no. 2, p. 024001, 2022.

[14] P. Bhoumik, C. Bailey, and K. Chakrabarty, "Defect-aware built-in self-test and dynamic repair for fan-out wafer-level packaging," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 34, no. 3, pp. 967–980, 2026.

[15] I. Jani *et al.*, "Characterization of fine pitch hybrid bonding pads using electrical misalignment test vehicle," in *IEEE 69th Electronic Components and Technology Conference (ECTC)*, 2019, pp. 1926–1932.

[16] Y. Kagawa *et al.*, "Impacts of misalignment on 1 μm pitch cu-cu hybrid bonding," in *IEEE International Interconnect Technology Conference (IITC)*, 2020, pp. 148–150.

[17] J. Jourdon *et al.*, "Search for copper diffusion at hybrid bonding interface through chemical and electrical characterizations," *Microelectronics Reliability*, vol. 126, p. 114217, 2021.

[18] S. H. Hahn *et al.*, "Contamination-free Cu/SiCN hybrid bonding process development for sub-μm pitch devices with enhanced bonding characteristics," in *IEEE 73rd Electronic Components and Technology Conference (ECTC)*, 2023, pp. 1390–1396.

[19] P. Bhoumik, C. Bailey, and K. Chakrabarty, "Defect analysis and built-in-self-test for chiplet interconnects in fan-out wafer-level packaging," in *IEEE 43rd VLSI Test Symposium (VTS)*, 2025, pp. 1–7.

[20] D. Das Sharma *et al.*, "High-performance, power-efficient three-dimensional system-in-package designs with universal chiplet interconnect express," *Nature Electronics*, vol. 7, no. 3, pp. 244–254, 2024.

[21] P.-Y. Chuang and E. J. Marinissen, "Chiplet interconnect repair for clustered defects with minimal propagation delay," in *IEEE Asian Test Symposium (ATS)*, 2025.

[22] T.-H. Wang *et al.*, "Test and repair improvements for UCIe," in *IEEE European Test Symposium (ETS)*, 2024, pp. 1–6.

[23] P.-Y. Chuang and E. J. Marinissen, "Chiplets' die-to-die interconnect repair language (IRL)," in *IEEE International Test Conference (ITC)*, 2025, pp. 37–44.

[24] W.-H. Lo, K. Chi, and T. Hwang, "Architecture of ring-based redundant TSV for clustered faults," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 12, pp. 3437–3449, 2016.

[25] L. Jiang, Q. Xu, and B. Eklow, "On effective through-silicon via repair for 3-D-stacked ICs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 32, no. 4, pp. 559–571, 2013.

[26] I. Lee, M. Cheong, and S. Kang, "Highly reliable redundant TSV architecture for clustered faults," *IEEE Transactions on Reliability*, vol. 68, no. 1, pp. 237–247, 2018.

[27] Q. Wang *et al.*, "A new cellular-based redundant TSV structure for clustered faults," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 2, pp. 458–467, 2018.

[28] T. Ni *et al.*, "LCHR-TSV: Novel low cost and highly repairable honeycomb-based tsv redundancy architecture for clustered faults," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 10, pp. 2938–2951, 2019.