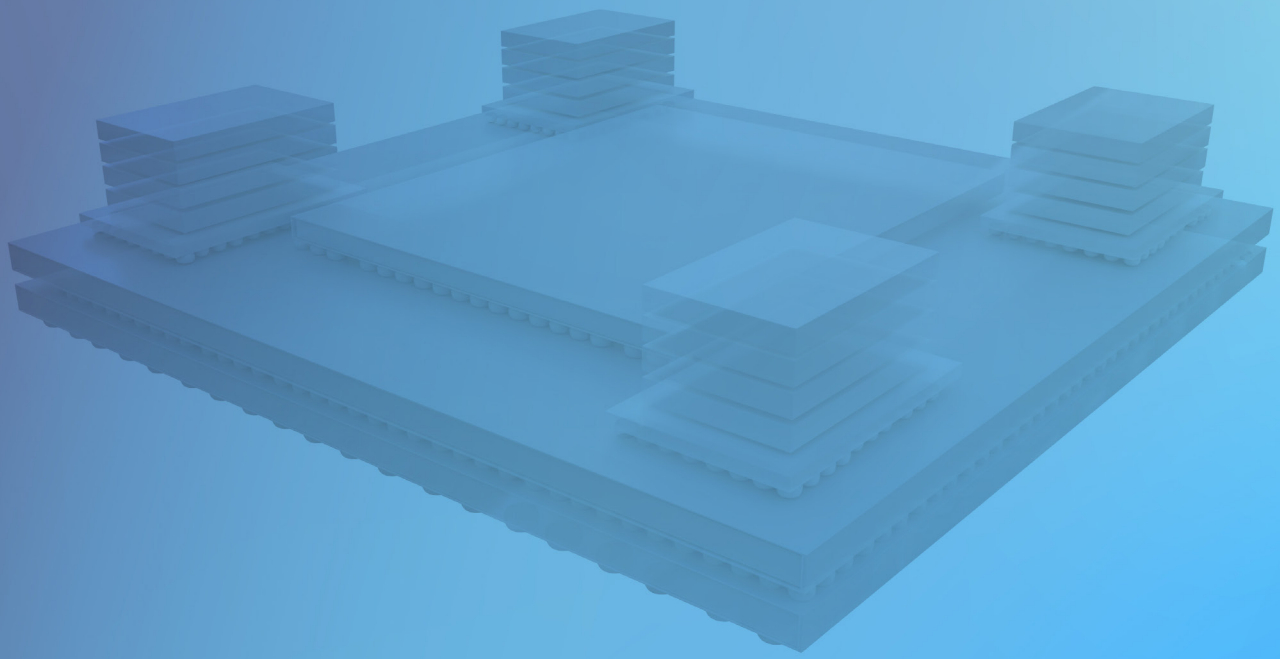
White Paper

Deep Data Analytics for High Bandwidth Memory (HBM) Reliability

In-field HBM monitoring and repair of GUC HBM subsystem using proteanTecs' Proteus™

Authors: Eyal Fayneh, proteanTecs; Igor Elkanovich, GUC



proteanTecs

GUC

Deep Data Analytics for High Bandwidth Memory (HBM) Reliability

Abstract

This white paper presents the use of proteanTecs' Proteus™ for HBM subsystem reliability based on deep data analytics and enhanced visibility, overcoming the limitations of advanced Heterogeneous packaging. It will describe the operation concept and provide results from a GUC 7nm HBM2E Testchip.

Introduction

High Bandwidth Memory (HBM) is a specialized form of stacked memory architecture that is integrated with processing units to increase speed while reducing latency, power, and size. It presents a premium DRAM offering for high-bandwidth applications such as next-generation supercomputers, graphics systems, and artificial intelligence (AI). HBM is rapidly evolving to meet the changing needs of the datacenter and networking industries and the technology has already gained significant adoption in the market, expected to grow at a CAGR of 32% by 2022[1]. HBM was adopted by JEDEC[2] as an industry standard in October 2013 and its second generation, HBM2, was accepted in January 2016.

HBM is comprised of DRAM stacks that are known as memory "cubes", each of which can have up to eight dies. HBM2 can support 8GB of memory per cube at a transfer rate of 256 GB/s. Four HBM cubes connected to a processor via an interposer will provide 32GB of memory with a bandwidth of 1TB/s. This improves system performance and increases energy efficiency, enhancing the overall effectiveness of data-intensive, high-volume applications that depend on machine learning, parallel computing and graphics rendering.

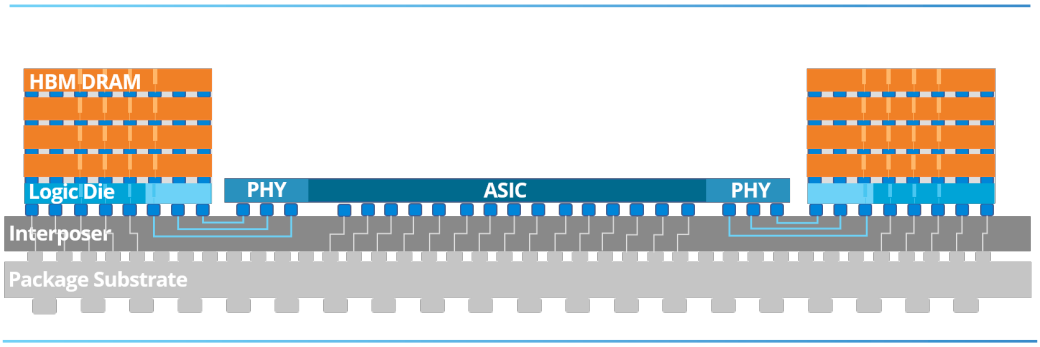


Figure 1: HBM Structure Connected to ASIC

An HBM subsystem is manufactured using advanced multi-die (heterogeneous) IC packaging, such as the TSMC CoWoS (Chip on Wafer on Substrate), and a typical 4xHBM2 subsystem has 13,600 micro-bumps that are used for signal connectivity. Visibility of HBM subsystems is limited by nature due to 3D integration technology and signal integrity problems are difficult to debug, validate and monitor. Multi-die HBM packaging introduces new reliability challenges which can lead to functional device failures in-field. These technologies are both inherently complex as well as expensive, therefore system failures afflict significant losses on manufacturers and service providers alike.

1. Market Reports World: High-bandwidth Memory Market 2019 Research
2. https://www.jedec.org/standards-documents/docs/jesd235a

Deep Data Analytics for High Bandwidth Memory (HBM) Reliability

This paper proposes an approach to in-field, in-mission monitoring of HBM subsystems using proteanTecs' Proteus™. Proteus is a one-stop software platform which applies analytics to data created by on-chip Agents™ (IPs), custom tailored to represent and automatically cover a specific design. The Agents monitor, detect and report from within chips, producing readouts which are inferred by machine learning algorithms and analytics tools to provide actionable insights and alerts on the electronic system's health and performance. These insights create a shared language of measurements across the value chain, providing feedback & feedforward correlation for unprecedented levels of quality assurance, failure prevention and resolution.

Agents™ embedded during design

Analytics & insights gained in product bring-up, high volume production and in-field

Figure 2: Proteus™ for Electronics Health and Performance Monitoring

Challenges of HBM Lifetime Reliability

A typical CoWoS chip has hundreds of thousands of micro-bumps (u-bumps). 3-8 u-bumps are used to route each of the signals or the power and ground. Due to the u-bump per signal redundancy, failure of a single u-bump may not necessarily affect the chip operation. However, an HBM PHY does not allow for redundancies due to the high-density routing, and one u-bump per signal is used for the entire HBM connectivity. In this case, a failure in any of the PHY or HBM u-bumps will lead to a chip operational failure. A typical 4xHBM2 includes 13,600 u-bumps for connectivity and the lack of redundancy creates a reliability challenge and risk of full HBM subsystem failure. At testing, the implications of a failed module incur significant monetary losses for manufacturers. In lifetime (field) operation, a failure in the HBM subsystem may affect the whole system and lead to an abrupt operational failure and unplanned downtime.

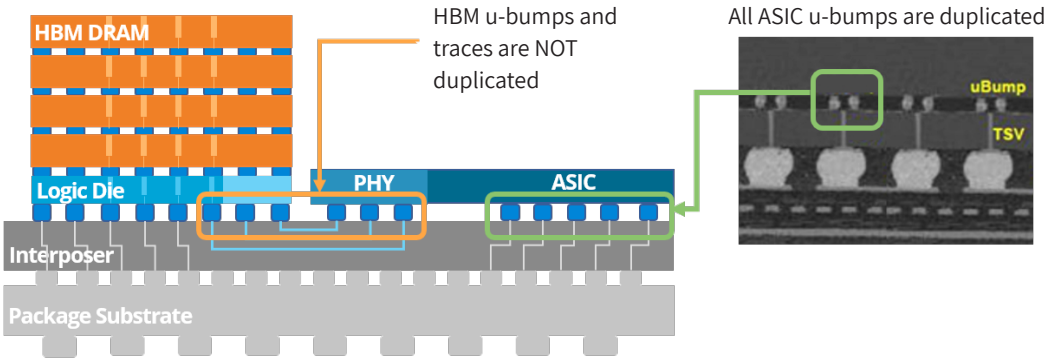


Figure 3: HBM PHY u-Bumps Lack Redundancy

Testing of the HBM subsystem is performed using industry standard detection tools, the first of which being a DC tool. It can detect open/short connectivity problems in a faulty lane and repair it using a “lane-repair” procedure, but lacks the ability to detect degradation over time in mission-mode. Another tool used is At -Speed BIST that can detect speed-related failures in a group of lanes. A group is defined by a D-Word or A-Word block which includes 32 data lanes; therefore, this method does not identify a specific failing lane. In any case, both tests are run offline so operation needs to be stopped and the HBM subsystem must be put in test-mode.

As part of the HBM subsystem testing, the eye-opening is measured and used to screen bad parts, but this method also lacks per-pin resolution. The eye-opening of the Data, Address and Control busses are measured at Write and Read modes. Each channel is separately measured and screened at a 32-bit data-bus resolution. The screening is only performed offline at boot operation and cannot be used in mission mode.

In case of failure detection, a lane-repair solution can be activated. There are two types of lane-repair operations: (1) hard lane-repair and (2) soft lane-repair. The hard lane-repair algorithm identifies failed lanes and replaces them with redundant lanes. The result is burnt using electrical fuses (eFuse) for further usage at next boot operation. A soft lane-repair routing can be performed as a temporary measure without burning the information into the eFuse. Both Hard and Soft repairs can be performed at ATE test and at boot operation in the field. Since monitoring HBM lanes requires activation in test mode, degradation monitoring in mission-mode is not covered.

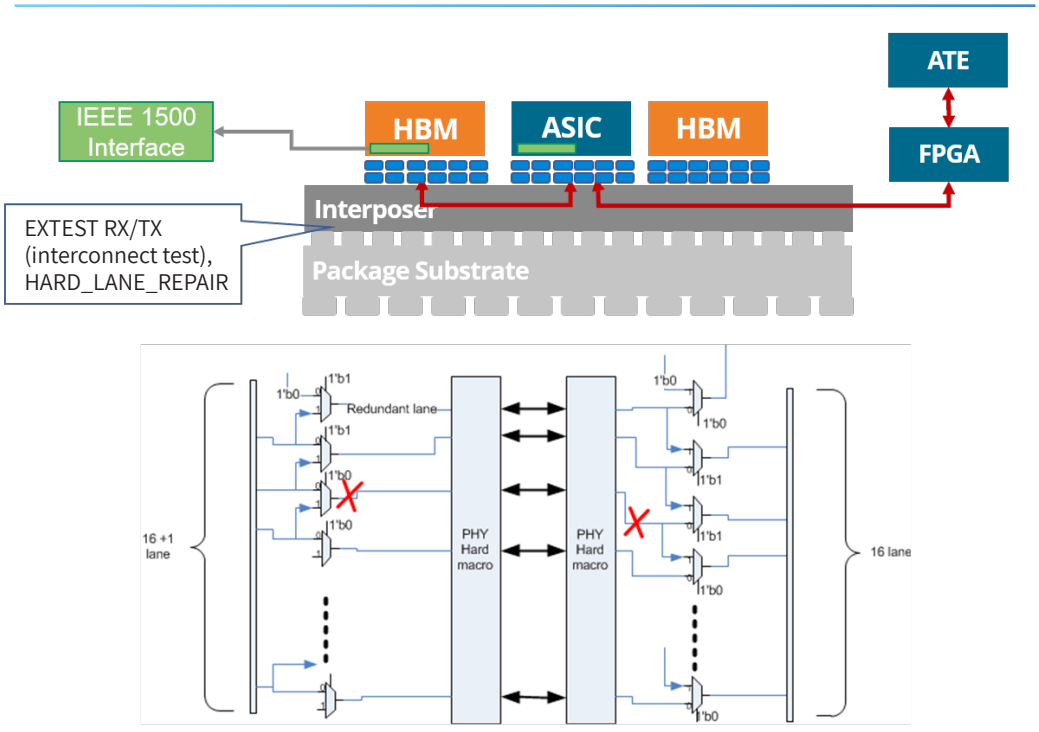


Figure 4: GUC Hard Lane-Repair Setup

HBM testing lacks parametric sensitivity so marginal lanes are not detected. These may lead to degradation over time and ultimately failure during lifetime operation.

A new method for achieving HBM visibility is needed to perform enhanced reliability monitoring, per pin and in-field. In additon, observability at system-level is required to enable fast bring-up and characterization.

Deep Data Analytics for High Bandwidth Memory (HBM) Reliability

Actionable Insights for HBM Reliability

proteanTecs' Proteus provides continuous visibility for enhanced HBM monitoring and repair, mitigating the limitations described in the previous section. Reliability monitoring is performed per pin and activated continuously in mission mode to detect degradation trends.

In this case study, Proteus was used in GUC's 7nm HBM2E testchip to provide actionable insights on the HBM subsystem's reliability. An HBM Agent was embedded into the chip to measure Near-End (NE) and Far-End (FE) signal integrity, which was extracted and uploaded to the Proteus analytics platform to provide actionable insights.

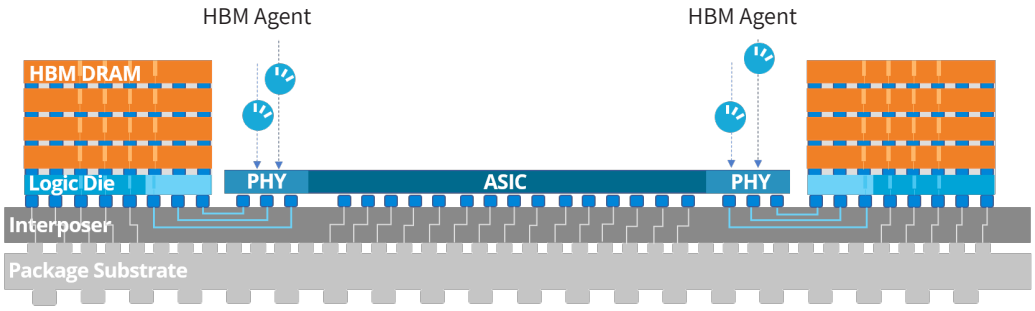


Figure 5: Proteus™ HBM Agent Integrated in GUC's PHY

The HBM Agent is comprised of I/O sensors per pin, that monitor Near-End (NE) and Far-End (FE) integrity insights, and a controller unit. The NE monitor represents Tx signal quality, derived from the ASIC driver strength, NE micro-bump integrity and interposer integrity. The FE monitor represents Rx signal quality, derived from the DRAM buffer driver strength, FE micro-bump integrity, interposer integrity and the ASIC's Rx buffer sensitivity. Both signal integrity insights are achieved by uploading the raw data created by the HBM Agent to the data analytics platform for inference.

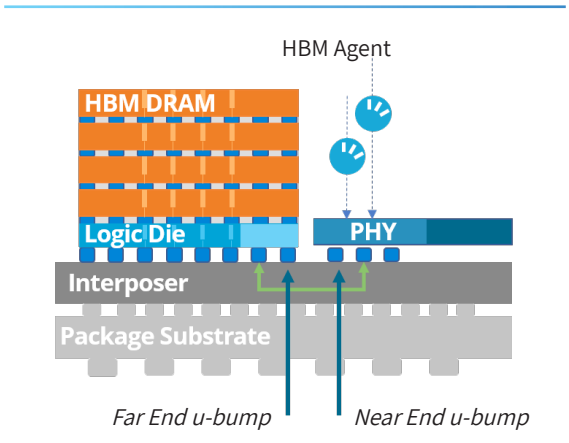


Figure 6: I/ O Sensors Monitor NE and FE Signal Integrity

Deep Data Analytics for High Bandwidth Memory (HBM) Reliability

Proteus was used to monitor degradation trends in order to predict system failure. By alerting on marginal performance of NE or FE signals, service providers can perform Predictive Maintenance, thus preventing HBM interface failure in mission-mode. Proteus identifies potential candidates for faulty-lane replacement, based on the signal integrity "insights" (a deep data measure for signal strength and quality correlated to signal slew rate and amplitude), and provides the information to the Lane Repair mechanism. The lane repair mechanism can be used to replace marginal lanes with redundant ones at scheduled maintenance cycles, preventing system failure due to signal quality degradation beyond margin limits.

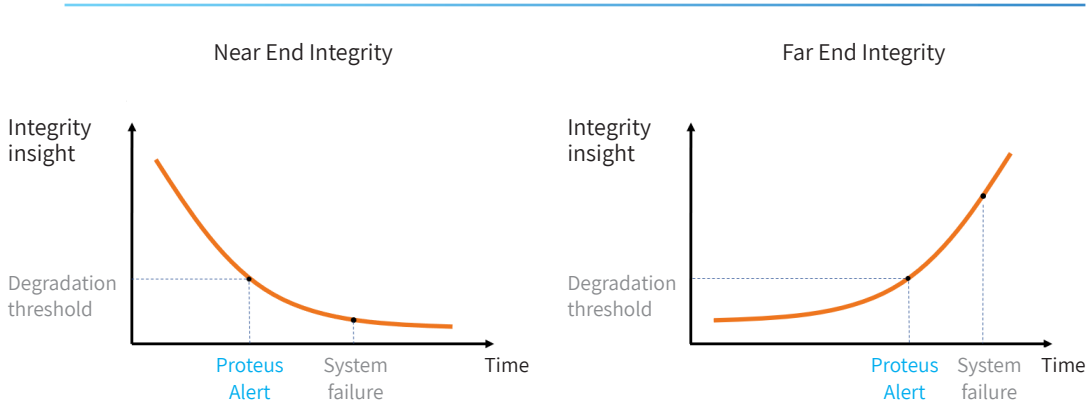


Figure 7: Degradation Monitoring and Alerts

At system bring-up and characterization, Proteus enables virtual probing of the signal amplitude and slew-rate for each pin, serving as an embedded "scope", without impacting the measured signal. This provides visibility of HBM signal parameters per pin during system characterization and validation, reducing time-to-market, achieving product optimization and increasing confidence in ramp-up.

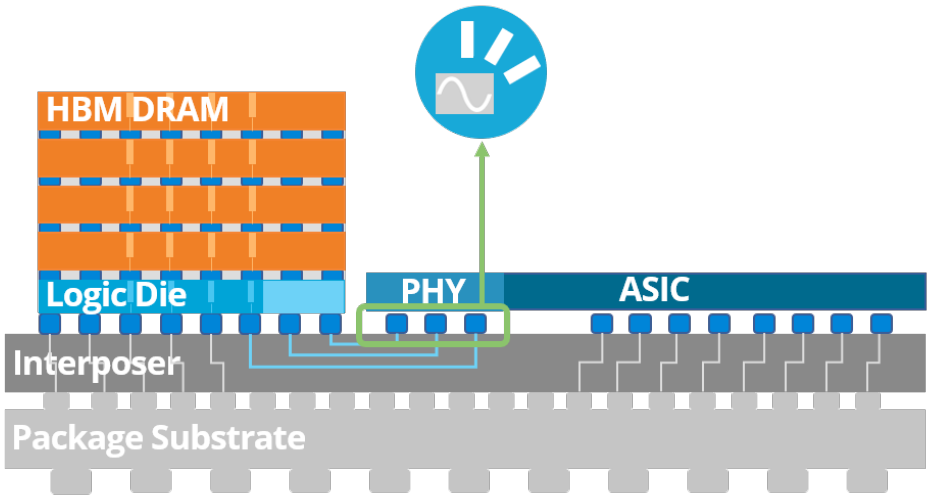


Figure 8: Proteus Embedded Virtual Scope

Deep Data Analytics for High Bandwidth Memory (HBM) Reliability

Implementation

Proteus was integrated into the following GUC testchips:

1. 7nm HBM2E 3.2 Gbps testchip EX0010A

i) This testchip contained two instances of 7nm testdies with GUC's HBM2E 3.2 Gbps PHY and Controller and HBM2/2E memory assembled on a silicon interposer.

ii) One of the 7nm HBM2E 3.2 Gbps testdies was connected to an HBM2/2E memory via straight interposer connection.

iii) An additional 7nm HBM2E 3.2 Gbps testdie was connected to the HBM2/2E memory via worst case long zig-zag interposer connection.

iv) EX0010A was assembled first with Samsung memory KHA884901X-MC13 (Aquabolt 2.4 Gbps speed grade). It was tested both at 2.4 Gbps and 3.2 Gbps using GUC's HBM2E 3.2 Gbps testdie with worst case long zig-zag interposer connection. The results in the next section were collected during this test.

v) EX0010A was also assembled with Samsung memory KHAA84901B-MC15 (Flashbolt 2.8 Gbps speed grade). It was tested with both worst (long zig-zag) and typical (straight) interposer connections. Testing results are available upon request.

2. 5nm HBM2E 3.2 Gbps testchip EY0003A

i) This testchip contains two instances of 5nm testdies with GUC's HBM2E 3.2 Gbps PHY and Controller and an HBM2E Flashbolt connected via worst (long zig-zag) and typical (straight) interposer connections. Testing is in progress as of April 2020 and results will be available in Q2 2020.

Results

Using the Proteus analytics software, the following insights were gained, demonstrating enhanced visibility for reliability monitoring and repair in GUC HBM subsystems. The results below were collected using the GUC EX0010A teschip with Samsung's memory Aquabolt, using worst case long zig-zag interposer connection. It was tested at both 2.4 Gbps and 3.2 Gbps.

I. Lane Degradation Monitoring and Repair in Mission-Mode Based on Near-End and Far-End Integrity Insights

Proteus provides a new method for correlating lane degradation to FE and NE insights, which are a measure of signal integrity based on inference of in-circuit measurements. The insights represent ASIC and DRAM driver strength, NE and FE micro-bump integrity, Rx sensitivity and interposer integrity. They enable tracking of lane degradation in mission-mode for failure prevention of systems in the field.

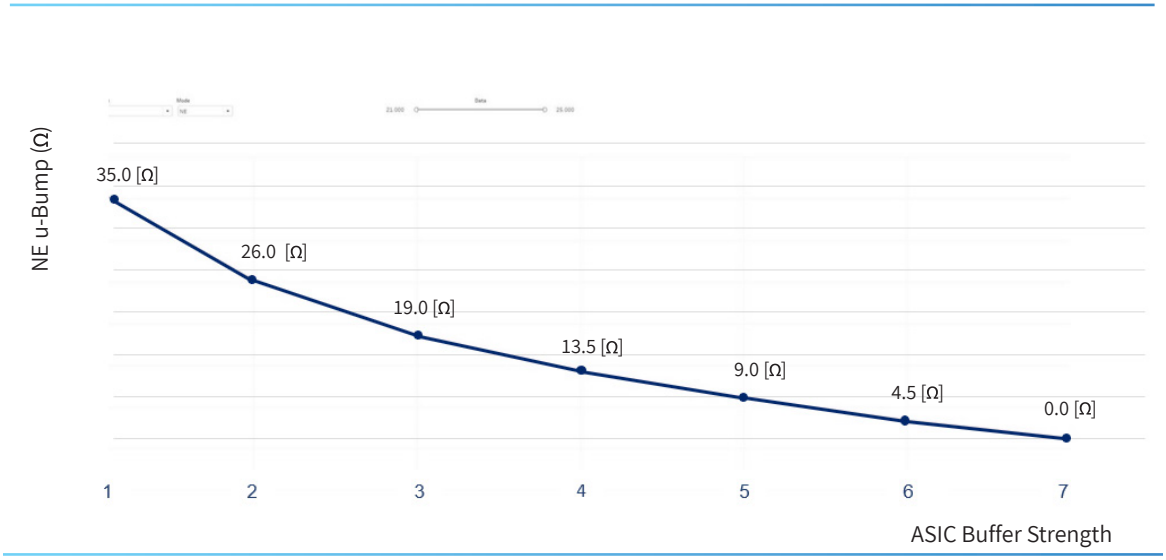


Figure 9: Near-End u-Bump Resistance Change Vs. ASIC Buffer Strength

Figure 9 shows the sensitivity of the NE insight vs. ASIC driver strength for a specific pin. Intentionally reducing the driver strength emulates the NE micro-bump resistance degradation. The sensitivity of the NE integrity insight is ~0.5 LSB which is translated to 3-6 Ω u-bump degradation (average of 4.5 Ω). Proteus alerts on degradation of the pin after 2 LSB (12-24 Ω degradation, still well within interface margins), therefore repair can be initiated before system failure.

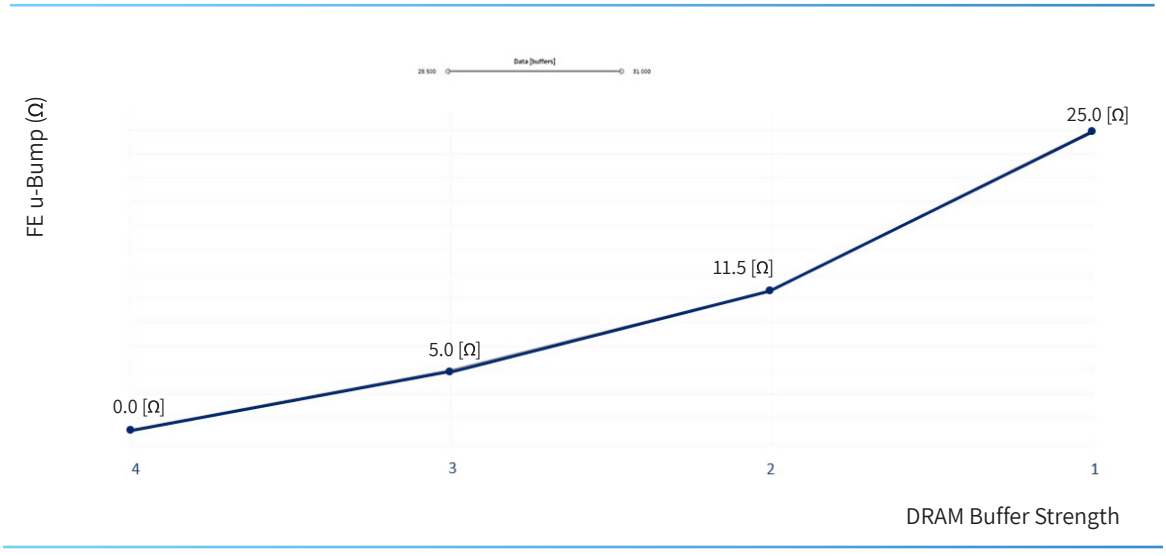


Figure 10: Far-End u-Bump Resistance Change Vs. DRAM Buffer Strength

Figure 10 shows the sensitivity of the FE insight vs. DRAM driver strength for a specific pin. Intentionally reducing the driver strength emulates the FE micro-bump resistance degradation. The sensitivity of the FE integrity insight is ~0.5 LSB which is translated to 4-6 Ω u-bump degradation (average of 5 Ω). Proteus alerts on degradation of the pin after 2 LSB (16-24 Ω degradation, still well within interface margins), therefore repair can be initiated before system failure.

I. Lane Degradation Monitoring and Repair in Mission-Mode Based on Rx Slew Rate

Proteus provides a method to measure signal slew rates (ps/V) at the Rx receiver pin. In mission-mode, a calibration cycle is conducted in which the slew rate is measured and correlated to the FE insights per pin. From that point on, only the FE insight is measured, without interference to system operation, and automatically correlated to the slew rate. This enables to track an electrical parameter for enhanced lane degradation monitoring in-field.

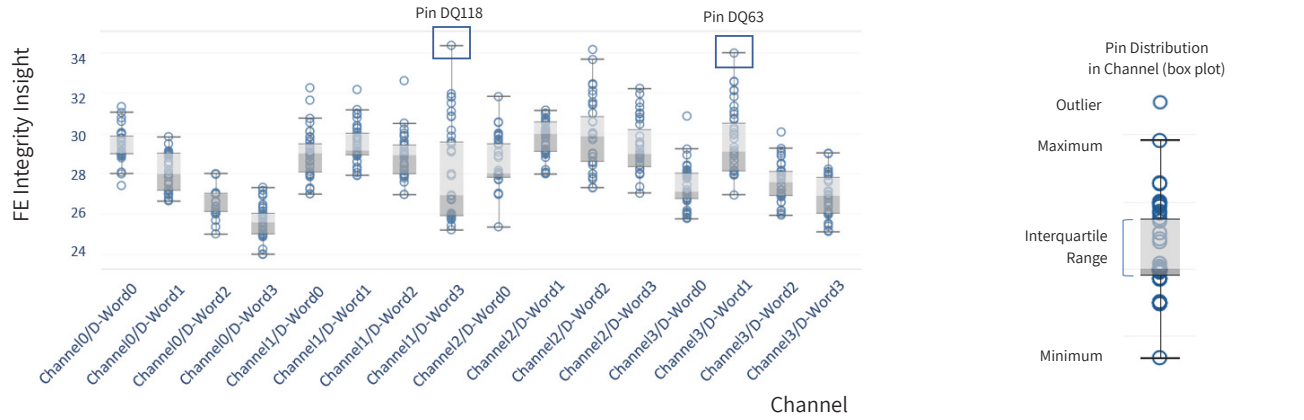


Figure 11: Far-End Integrity Insight per Channel and Pins in Channel

Figure 11 shows the distribution of the FE integrity insights, measured at DRAM driver strength 4 (maximum strength). Lower insight values demonstrate stronger signal integrity and higher insight values demonstrate weaker signal integrity (e.g. Channel0/D-Word3 has stronger signal integrity than Channel0/D-Word0). The FE insights monitor DRAM buffer driver strength, FE micro-bump integrity, interposer integrity and the ASIC's Rx buffer sensitivity and are correlated to the signal slew rate and amplitude. The insight was measured per channel, D-Word block and pin, and reflects insights at time-zero (no lifetime degradation). Two observations were gained: [1] Greater variation of the FE insight is observed in certain blocks (i.e. Channel1/D-Word3 has greater variation than Channel0/D-Word3). This is a result of data-dependent ISI, derived from varying data patterns transmitted, which lead to different amplitudes (such as a lone bit). [2] Furthermore, deviation of certain pins from their respectful D-Word block and from the distribution of all channels is observed, pointing to weakness from a statistical point of view (e.g. pins DQ118 and DQ63). As observed in **Fig. 13** below, the Rx Slew Rate of the same deviant pins (e.g. DQ118 and DQ63) was higher than their respective D-Word block and higher than the distribution of all channels, indicating parametric marginality.

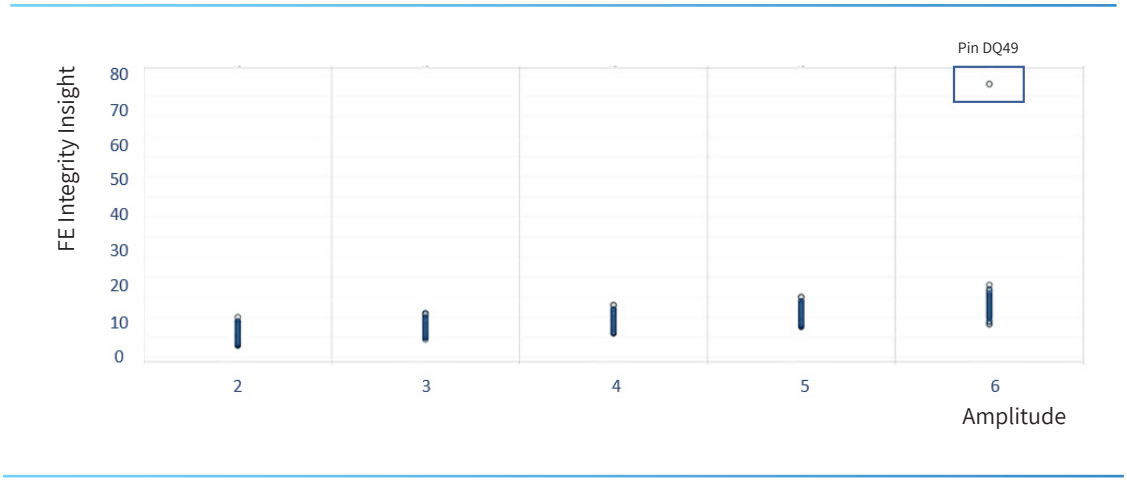


Figure 12: Rx Signal Amplitude at Pin

Figure 12 shows FE signal integrity using an embedded virtual scope that measures signal amplitude and slew-rate for each pin. Deviant values are caused by weak Rx signal at ASIC pin (e.g. DQ49).

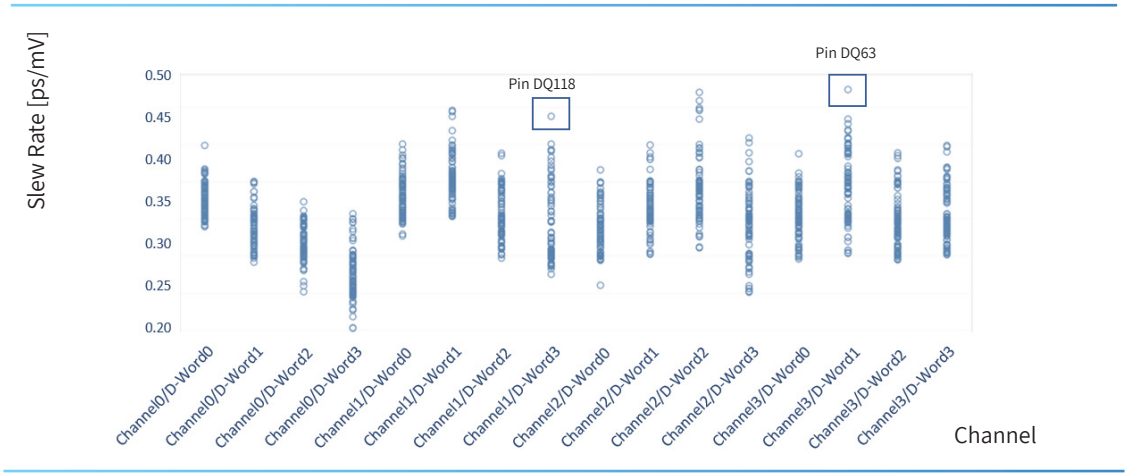


Figure 13: FE Signal Slew Rate per Channel and Pins in Channel

Figure 13 shows the distribution of the Rx slew rate, measured at DRAM driver strength 4 (maximum strength), per channel, D-Word block and pin. It reflects insights at time-zero (no lifetime degradation). Deviant pins are observed (e.g. DQ118 and DQ63), demonstrating measured slew rates higher than their respective D-Word block and higher than the distribution of all channels, indicating parametric marginality. The same pins also demonstrated higher FE Integrity Insights, pointing to weakness also from a statistical point of view (as observed in **Fig. 11**).

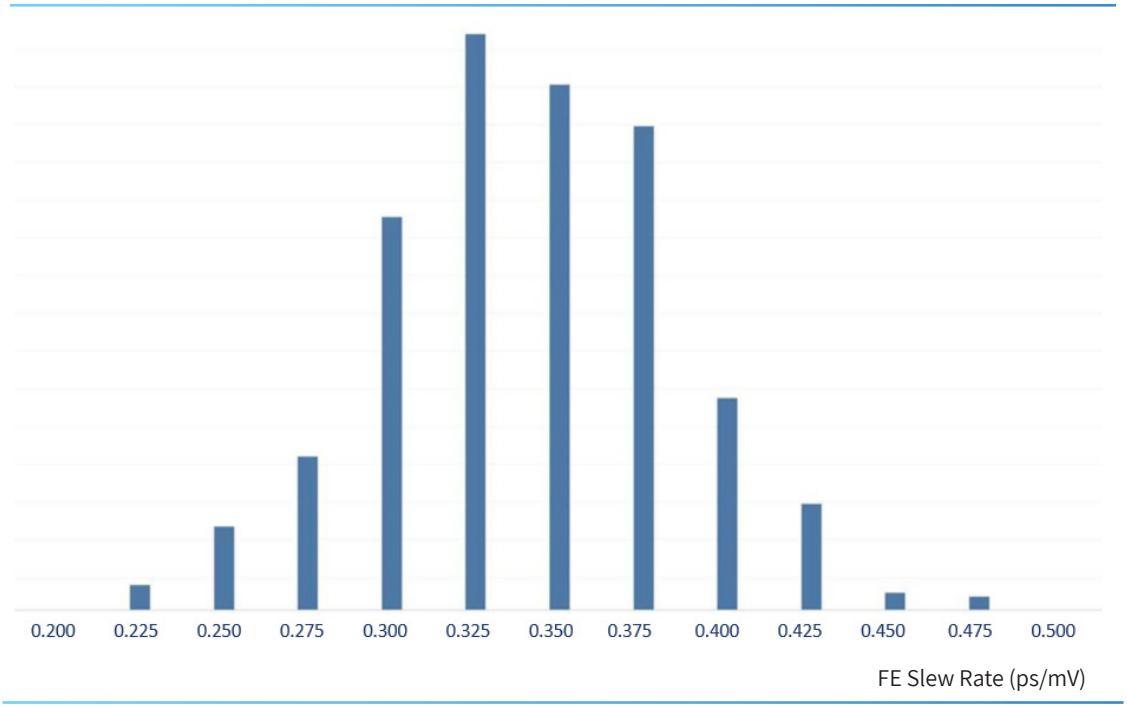


Figure 14: FE Signal Slew Rate Distribution for all Groups and all Pins

Figure 14 shows the statistical analysis of the signal slew rate per all groups and all pins. The measurement sensitivity equals 10ps/50mV. Proteus measures the slew rate per pin at the calibration cycle, and the values are then correlated to the FE insight **(**as shown in **Fig. 15).** This is used to monitor the Rx performance in mission mode and alert on degradation before reaching system failure.

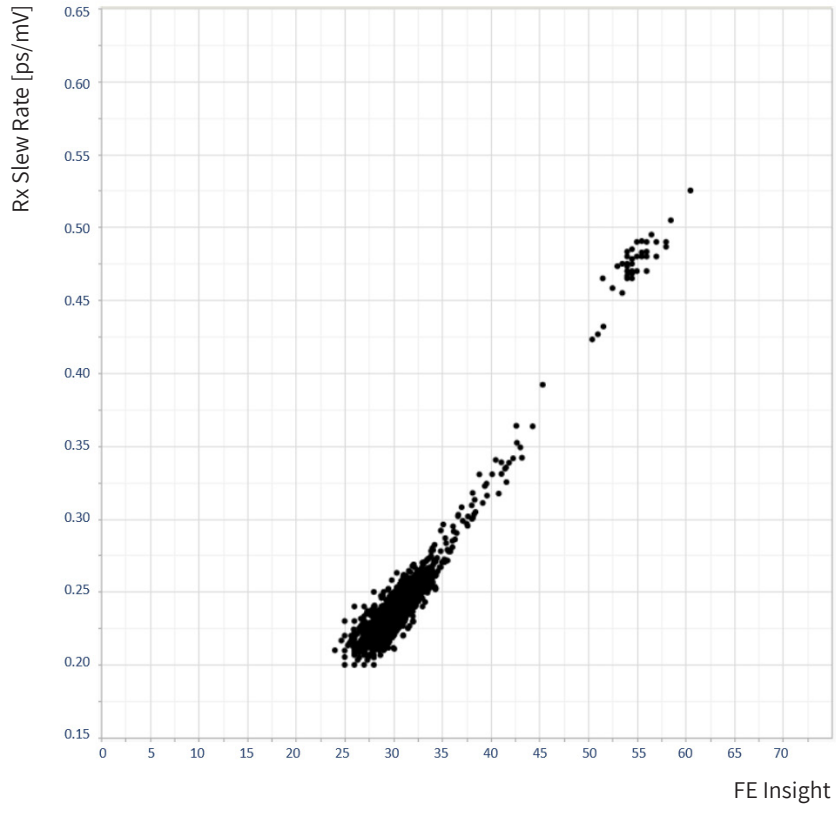


Figure 15: Correlation of Rx Signal Slew Rate to FE Insight

Conclusion

As the complexity of heterogeneous packaging continues to develop, new reliability challenges arise. An emerging approach to HBM subsystem monitoring and repair is introduced, for advanced in-field reliability assurance. By applying machine learning algorithms and analytics to data created by on-chip Agents (IPs), proteanTecs' Proteus provides actionable insights and alerts on the system's lifetime operation. As observed in GUC's 7nm HBM2E testchip, continuous monitoring of signal degradation is performed, per-pin and in-mission. Service providers gain the visibility they need to perform Predictive Maintenance, detecting and repairing faults before they become system failures.

List of Figures

proteanTecs Ltd.
36 Kdoshei Bagdad Dr.
Haifa, Israel 33032
www.proteanTecs.com

Global Unichip Corporation
No. 10, Li-Hsin 6th Road, Hsinchu Science Park
Hsinchu City 30078, Taiwan
www.guc-asic.com

proteanTecs

GUC