Embedding SRAM chiplets in fan-out interposer for memory disaggregation

Dany-Sebastien Ly-Gagnon, RNB-2-105, dany-sebastien.ly-gagnon@intel.com

Jiajing Wang, SC12-442, jiajing.wang@intel.com

Ping-Chen Liu, SC9, ping-chen.liu@intel.com

DerChang Kau, RNB-2-102, derchang.kau@intel.com

1 Introduction

In recent years, new technology nodes have enabled logic to continue its area scaling trajectory, but SRAM has been facing difficulties in continuing the pace in scaling. With wafer costs increasing on latest technology nodes, the case to disaggregate the large memory arrays into SRAM chiplet companion dies from mature technology nodes is becoming more compelling, as it allows SRAM chiplets to be fabricated on less expensive wafers, while logic can continue to leverage the benefits of technology scaling. For memory disaggregation, a 3D vertical interconnect (e.g. vertical stacking) is preferable over side-by-side connectivity, since it provides a larger area for wide-IO interconnects and reduced trace length, which can improve memory bandwidth, latency and power.

Several methods can be used to enable 3D integration of SRAM chiplets near the logic die. Notably, face-to-back hybrid bonding (HBI) has been used by AMD to stack an SRAM chiplet on a logic die[1]. While HBI can provide higher interconnect densities with pitch of 9um or below, the use of HBI requires tight control of the bonding interfaces and may be difficult to achieve when combining technologies from different foundries, thus limiting technology options. Large silicon interposers (such as Foveros[2] or CoWoS-S[3]) could also be used as active silicon to integrate large SRAM arrays and provide a 3D integration, but this approach is not cost effective, since the memory arrays are unlikely to fill the entire interposer area, and thus lead to a poor silicon area utilization, which increases the overall cost.

Organic fan-out interposers[4,5], on the other hand, have recently been used in larger chiplet systems in conjunction with passive embedded bridges for side-by-side connectivity. Such platform could be used to embed SRAM chiplets to enable vertical stacking of an SRAM chiplet. This would enable the selection of a technology node for SRAM with the best figure-of-merit (FOM) for the intended application (see Figure 1a), to a platform that is already available in a range of capabilities (see Figure 1b).

This platform would provide several advantages for memory disaggregation. First, it enables the embedded SRAM chiplet to be right-sized, and optimized for yield and reticle field utilization, without penalty, since its design can be decoupled from any requirements imposed by the logic die, which is likely to use a more advanced and more expensive technology node. Second, it enables access to a wider range of technology nodes, which can be provided by either internal or external foundries, independently of the logic dies. Furthermore, it provides a path for modularity, such that chiplets can be refreshed or reused as required by the product cycle, enabling a richer set of product skews, which can be defined by the presence or absence of specific chiplets embedded in the interposer.

In this paper, we present the design and architecture of an SRAM memory chiplet design with an enhanced wide-IO interface that can be embedded in a fan-out interposer, enabling a vertical face-to-face 3D interconnect to a host logic die through microbumps. We implemented a 2MB SRAM chiplet with a 576-bit wide read/write access through a 25um microbump pitch interface on a mainstream advanced logic technology node in high-volume manufacturing to serve as a case study to evaluate bandwidth, latency, power and area design trade-offs. We describe the implementation that we adopted to enable an efficient area utilization for the die-to-die interconnects and the SRAM memory arrays. This design can achieve a read/write bandwidth in full-duplex of 102.4GB/s, across a 1152 microbump interface with read / write latencies of 4.375 / 2.5 ns, respectively. In this case study, we find that while the active power consumed under full workload by the memory (including leakage) accounts for ~35% of the total power consumption, while an aggressive reduction in microbump pitch may only lead to ~25% of reduction in power.

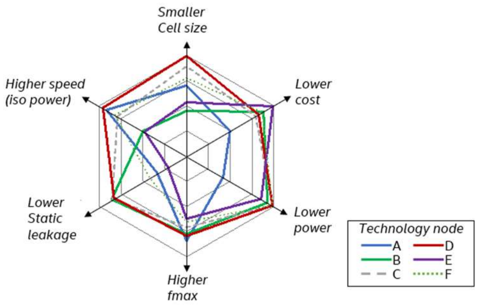

Figure 1a: FOM for tech nodes across silicon foundries

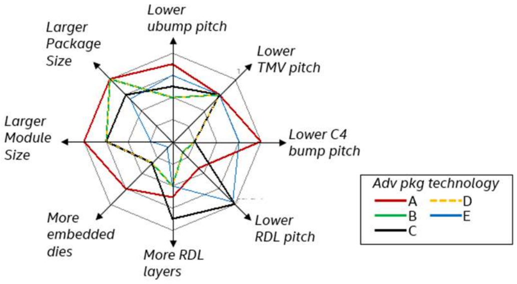

Figure 1*b: FOM for organic fan-out interposers across OSAT*

2 Description

2,1 System architecture concept

An example system is shown in Figure 2, where a compute chiplet and peripheral chiplet are interconnected through an embedded bridge chiplet, and a memory chiplet resides below the compute chiplet. In this configuration, the compute chiplet is interconnected to the memory chiplet through a vertical microbump interconnection layer. This type of advanced package can support optional RDL layers above or below the embedded die, to increase flexibility in routing signals and power tracks in-and-out of the package. This example can be further scaled to include additional bridge dies, memory chiplets or other active chiplets in the embedded dies layer to enable a wider range of applications.

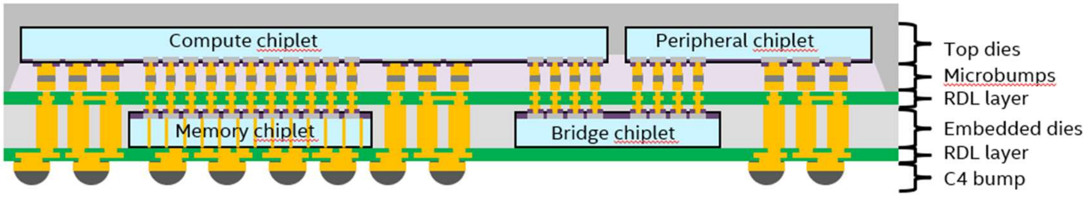

Figure 2: Example cross section of an embedded memory chiplet embedded in a fan-out interposer

2.2 SRAM chiplet architecture concept

Figure 3 shows a generic SRAM chiplet architecture. This chiplet has “*c*” channels, with each channel having independent IO to access their individual internal memory banks. To increase bandwidth and reduce requirements on individual memory macro read/write completion time, multiple banks can be accessed in parallel and time multiplexed through the global IO. As such, given a channel with “*b*” banks, the effective macro read/write completion time is reduced by a factor of “*b*”. Each bank contains *m* x *n* SRAM memory macros. Each read / write operation will select “*m*” of the *m* x *n* banks per read/write operation. The total number of memory macros (*m* x *n*) can be determined by optimizing for total memory size, the read/write completion time, bus width and area requirements for a specific application.

As an example, let’s consider a SRAM memory chiplet of 64MB, composed of 8 independent channels, each with 8MB capacity. With each channel using 2 banks of 4MB each, the read/write completion time for a memory macro can be relaxed by a factor 2 as the global IO can interleave access between each of the banks. In this scenario, small SRAM memory macros with 4k words and 128 bits can be used in a *m* x *n* = 4 x 16 configuration, to provide an overall memory capacity of 4MB and 512-bit wide interface. Assuming each bank interface is using an effective 800MHz data rate, and each bank using 512 bits on the read/write interface, the global IO on each channel would present a data rate of 102.4 GB/s. The aggregated overall bandwidth across all channels would total 819.2 GB/s to access an overall 64MB capacity.

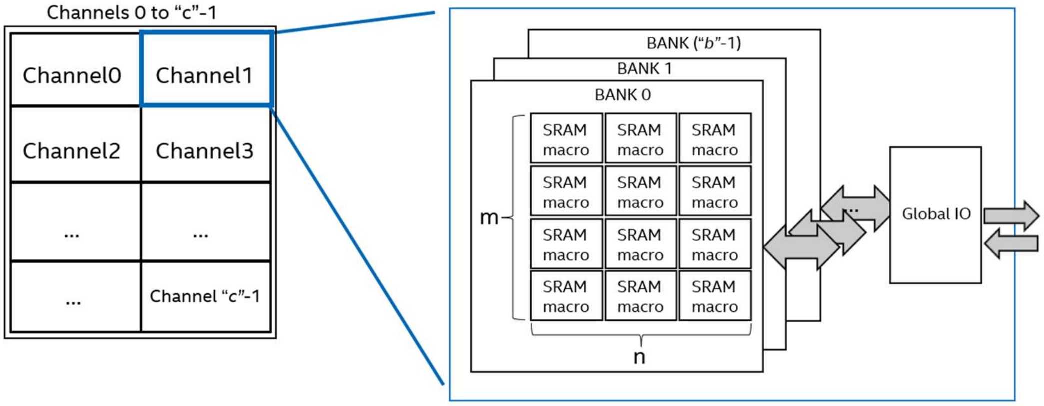

Figure 3: Block diagram of a memory chiplet

3 Results & Discussion

3.1 Case study for a 2MB channel with 102.4GB/s interface

To evaluate the bandwidth, latency and power of a SRAM memory chiplet, we implemented a single channel with an interconnect layer that uses a microbump pitch of 25um. For this case study, we consider a 2MB channel with a 576-bit wide read/write interface, consisting of 2 banks of 1MB, each with 4 x 4 SRAM macros. Each of the SRAM macro is configured as a 4k word x 144 bits, to include an 8:1 bit-level redundancy and lead to an effective 0.125MB capacity. To maximize bit density, high-density 6T SRAM cells are used in the memory macros. The SRAM macro configuration was targeted to operate at an 800MHz frequency for read/write operation and optimized for area and power. This configuration provides an effective read/write bandwidth of 102.4Gbps at the die-to-die interface.

The floorplan shown in Figure 4 was implemented on an advanced logic technology commercially available in high volume production. A die photograph of the implemented structure is shown in Figure 5. The microbumps are arranged in a hex lattice at 25um pitch and overlap with the SRAM memory array. There are 1152 microbumps for the read and write datapath, providing a full-duplex die-to-die interconnect between the dies. An additional 16 microbumps for address & control, 1 microbump for the clock feed-forward. For this implementation, two additional microbumps on each column of microbump are added to provide microbump redundancy. These redundant microbumps provide opportunity to repair up to two defects per column, and provide an additional layer of redundancy. The clock feed-forward microbumps are positioned in the center of the array, to facilitate the clock tree synthesis. The area between the two SRAM banks was also used to incorporate DFT.

In this implementation of a 2MB channel with 576-bit wide read/write interface, the overall microbump area is comparable to the frontend area required by the SRAM memory macros, the drivers and the global IO blocks. It is interesting to note that because the area of this design is mainly limited by the frontend (e.g. SRAM memory macro and driver size), and a reduction of the microbump pitch may not lead to a significant area reduction.

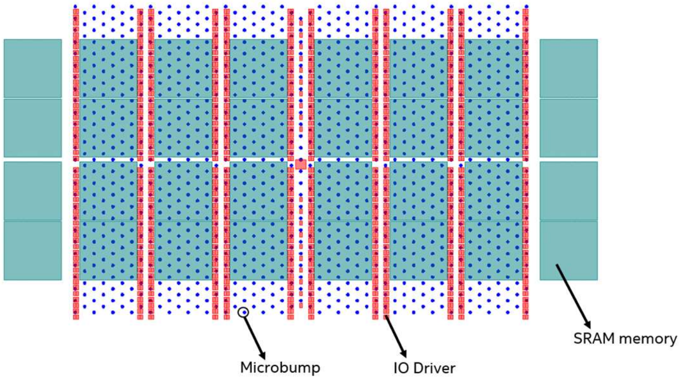

Figure 4: Floorplan for sample 2MB channel

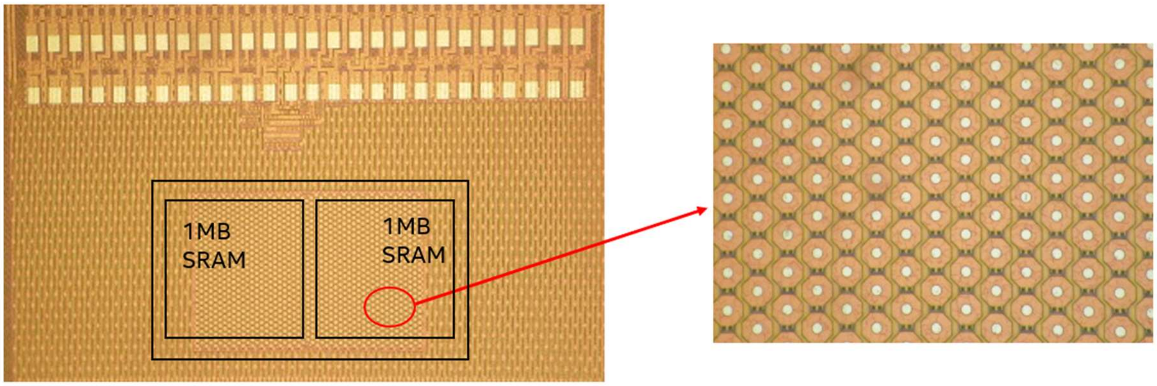

Figure 5: Die photograph of the implementation, with close-up on the die-to-die interface

For the die-to-die interface, the IO drivers are positioned nearby their associated microbump, as shown in Figure 6. Since the microbumps overlap with the SRAM memory macro area, the IO driver frontend was placed in the trenches between the SRAM arrays to minimize routing distance and parasitics. In this design, we established 4 unique different routing lengths between the drivers and the microbumps (denoted as D0, D1, D2, D3 in Figure 6). These 4 different routing lengths form a group, which is repeatedly tiled to enable the enhanced wide-IO connectivity. Because there are only 4 distinct routing paths, the driver sizing can be done according to the effective routing length and associated parasitic in order to meet timing closure.

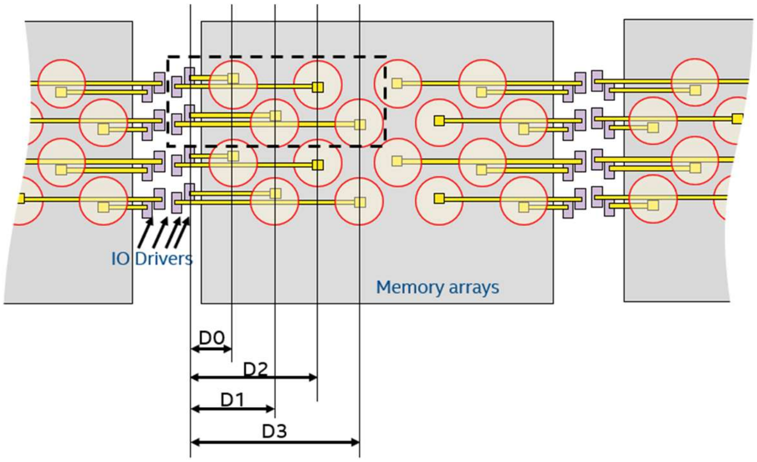

Figure 6: Floorplan overview of the die-to-die interconnect interface

3.2 Datapath

A simplified schematic of the datapath is shown in Figure 7. There are three main path segments in the datapath, namely the die-to-die interconnect between the TX and RX IO drivers, the path segment between the drivers and the global IO, and the path segment between the global IO and the memory. Registers were inserted in the datapath between each path segment to reduce timing closure risk in this design. The clock is fed forward from the top die logic chiplet down to the SRAM chiplet. The clock design is also an important challenge in this system and will be discussed elsewhere.

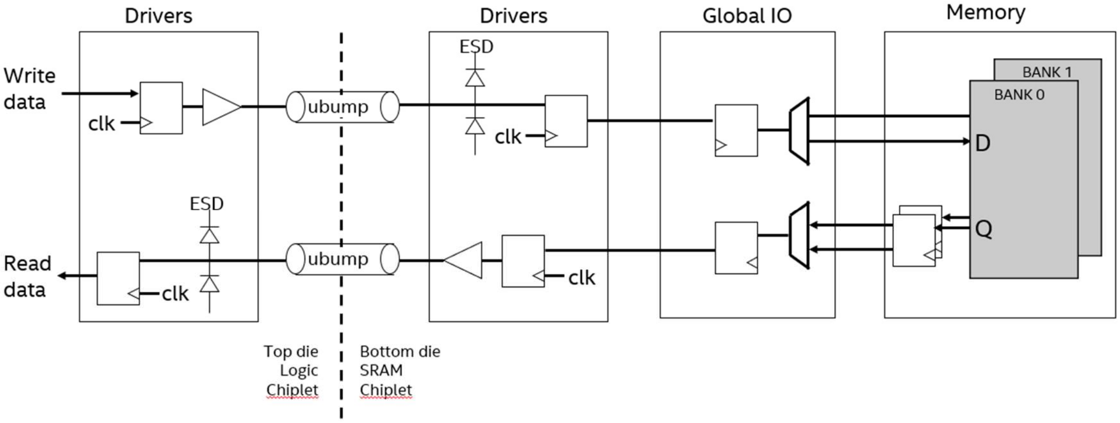

Figure 7: Simplified schematic of the datapath

3.3 Latency and Power consumption

In this design, the total path latency for a read operation is 7 cycles from pad-to-pad, which translates to a latency of 4.375ns for a base clock of 1.6GHz. For a write operation, the total path latency is 4 cycles, or 2.5ns.

The total power consumed during a 50/50 read/write workload running at a full bandwidth of 102.4GB/s was extracted from simulation on this 2MB channel design. Figure 8 summarizes the effective power used during the workload. The power used for die-to-die interconnect accounts for less than 20% of the total power. The power consumed in the driver-to-global-IO is around 22%, while the power consumed in global-IO-to-SRAM is around 25%. The active power used by the memory itself accounts for less than 20%, while leakage accounts for around 12%.

A further reduction of microbump pitch may enable floorplan improvements to reduce latency and power consumption. As an example, we expect a 15um pitch, for example, to allow a repositioning of the drivers and enable latency to reduce down to 3.125ns / 1.875ns for read / write, respectively. In addition, the power consumption could be reduced by approximately 20-25%, by reducing the power required for datapath routing in the driver-to-global-IO path segment. However, the power consumed by the global-IO-to-SRAM and the memory are technology dependent and do not depend on microbump pitch.

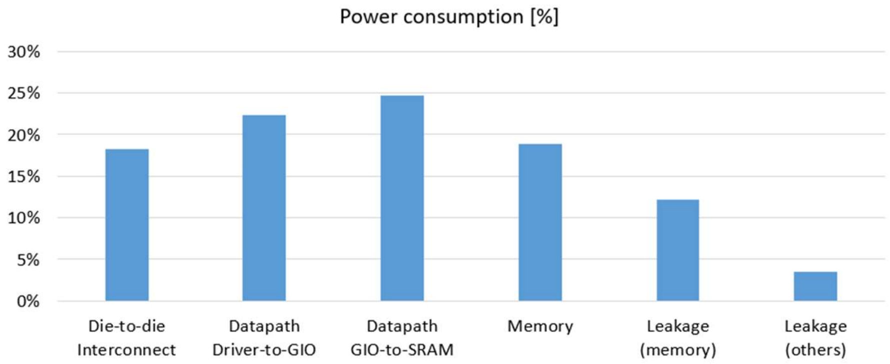

Figure 8: Power consumption breakdown.

3.4 Die-to-die interconnect

In this section, we take a closer look at the die-to-die interconnect path. A sample path for die-to-die interconnect is shown in Figure 9a and Figure 9b, for D0 and D3 drivers, respectively. To simplify analysis, the interconnect is assumed to be symmetrical with the top die logic side. As the IO drivers co-located between the SRAM memory arrays, upper metal routing layers provide access to the microbump. In the D0 driver case, the microbumps located directly above the IO driver. In the D3 driver case scenario, the interconnect requires requires longer routing in the upper metal layers, which may lead to an increased parasitic capacitance in comparison with D0 drivers.

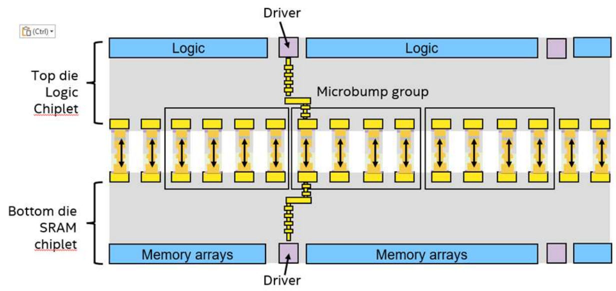

Figure 9a: Die-to-die interconnect for D0 driver

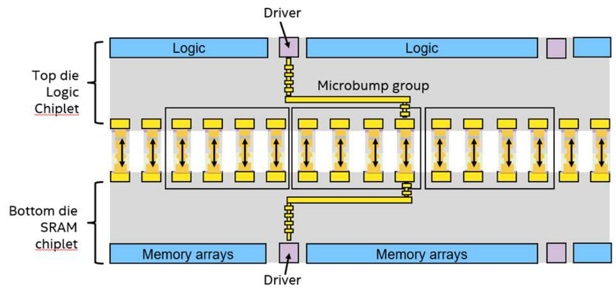

Figure 9b: Die-to-die interconnect for D3 driver

We extracted the parasitic capacitance for each of the routing lengths (D0, D1, D2 and D3) in this design. The total capacitance of the die-to-die link is shown in Figure 10, normalized to the die-to-die total capacitance on D0 driver, along with the contributions from the drivers, the upper metal routing and the microbump parasitic capacitance. The contribution of the drivers includes both the TX and RX driver (and ESD device in the drivers). The overall contribution of the microbump itself is relatively low, given that the fill material between the dies is of low refractive index, and this link does not include any TSV since the two chiplets are connected face-to-face.

It is important to note that while the capacitance increases for interconnect links with longer routing distance, in this design, we see the total capacitance only increases by up to 20%, and the largest contribution is attributed to the TX and RX drivers. Also given the even distribution between D0, D1, D2 and D3 drivers, the overall impact of routing is amortized on the d2d interconnect, and the additional parasitic capacitance due to the microbumps not being directly located above the drivers is only 10%.

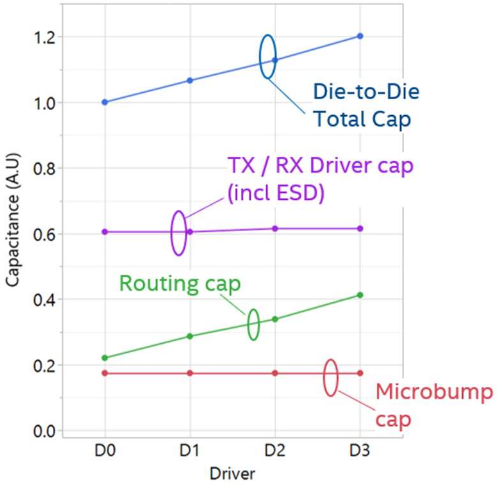

Figure 10: Total capacitance in die-to-die and contributions from driver, upper metal routing and microbump

4 Conclusion and Future Plans / Use

The 2MB SRAM chiplet case study as presented here was taped-in in August 2023 for silicon validation and silicon wafers became available at the time that this manuscript in redaction. The SRAM chiplet will be integrated as an embedded active chiplet in an organic fan-out interposer and validation of this proof-of-concept test vehicle will take place in coming months.

As the wider adoption of advanced packages provides increased opportunity for heterogenous integration of silicon dies from different foundries, we expect organic fan-out packages to pave the way towards low cost integration of large L3 or last-level cache memory with an enhanced wide-IO interface. The vertical face-to-face 3D stacking of chiplets with microbump at pitch of 25um or below can provide an interconnect density that can be well utilized to enable high bandwidth, low power, low latency and area efficient designs.

5 Acknowledgments

We would like to thank the Muddy Creek team for the design enablement and test vehicle tape-out, including Bryant Chang, Kt Chen, Yen-Jen Chen, Kurt Chu, Eddie Flores, Soumya Jayaraman, Kt Kuo, Ping-Chen Liu, Knuth Lu, Vijaya Maganti, Shibu Menon, Chung-Ching Peng, Zih-Nan Tseng, Jiajing Wang, Ch Yang, Peter Yeh and Wei Zhou.

6 References

[1] J. Wuu et al., 2022 ISSCC, pp. 428-429, DOI:10.1109/ISSCC42614.2022.9731565

[2] D. B. Ingerly et al., 2019 IEDM, pp.19.6.1-19.6.4, DOI:10.1109/IEDM19573.2019.8993637

[3] P. K. Huang et al., 2021 ECTC, pp.101-104, DOI:10.1109/ECTC32696.2021.00028

[4] M. L. Lin et al., 2022 ECTC, pp. 1-6, DOI:10.1109/ECTC51906.2022.00008

[5] J. Lin et al., 2020 ECTC, pp. 14-18, DOI:10.1109/ECTC32862.2020.00015