Advanced Packaging Drivers/Opportunities to Support Emerging Artificial Intelligence Applications

Luke England, Eric Tremble, Igor Arsovski, Wolfgang Sauter
AveraSemi, Burlington, VT, USA luke.england@averasemi.com

Abstract

The adoption of new artificial intelligence products for data center and edge computing applications drives a need for advanced packaging technologies. Solutions based on 3D/TSV, fine pitch interconnects, heterogeneous packaging, and scalable computing are prime examples. These packaging technologies will be described along with the benefits they offer. In addition, interface design considerations for die-to-die connections will be discussed.

(Keywords: Advanced Packaging, 3D, TSV, Fanout, Artificial Intelligence)

Introduction

The rise of artificial intelligence (AI) toward mainstream usage has been recently supported by the power and computing performance increases allowed by advanced node logic. Initially performed using CPUs due to their high level of flexibility, AI processing has made the transition through GPUs and FPGAs for higher performance, to fully differentiated ASIC chips designed to perform a specific purpose. Although these advanced node ASIC products provide a much improved power to performance ratio, power consumption is still a major concern. As an example, data centers consume roughly 10% of electricity generated worldwide in 2018 [1], and this growing trend is likely to continue. Since more energy is used to move data across multiple chips in a system than actually computing, it is imperative to bring chips as close together as possible to minimize the interconnect length. New advanced packaging and heterogeneous integration techniques are well positioned to play a key part in future power and performance improvements. This paper will review 3D packaging drivers, and solutions that will support the proliferation of AI applications in this period of "More Than Moore" scaling.

3D Packaging

A. Memory Proximity

A typical artificial intelligence functional schematic is shown in Fig 1a [2]. In addition to a logic unit and processing elements (i.e. MACs), the AI ASIC also contains a large local memory buffer (i.e. cache) to reduce the number of off-chip memory transactions to a large external memory (typically DRAM). When compared to the energy needed to compute in the Processing Element (PE), the energy cost of moving data to DRAM and back is 200X higher, as illustrated in Fig 2 [2]. Much of this energy penalty is due to long wire lengths between the ASIC and DRAM. Although the usage of 2.5D packaging with Si interposers brings DRAM closer to the ASIC, the need for higher memory bandwidth is pushing DRAM power as a major system level contributor. One way to address the problem is through 3D packaging to further reduce the wiring distance between the ASIC and DRAM. As an example, moving an HBM stack from a proximity of ~5mm on a Si interposer to ~50um directly stacked on the backside of the ASIC (in terms of wiring length) effectively eliminates the capacitive load, which reduces the capacitive switching power by 97% (or roughly 3W of the total HBM power). Further bandwidth and power efficiency improvements can be achieved with interfaces specifically designed for 3D integration, such as 3D SRAM. Table 1 shows the progression of power savings with memory proximity to the ASIC for various memory types.

B. SRAM Partitioning

3D integration can not only help at the system level, but can also be used to optimize chip partitioning. A typical ASIC or processor architecture contains a small number of large compute cores with shared blocks of cache. As AI applications are further developed, new architectures will be required to support them. For example, partitioning the SRAM cache and placing dedicated SRAM dies off-chip in a 3D configuration directly above the processing cores provides several significant advantages for large ASIC configurations:

- Large quantities of cache available for each processing core.
- Power reduction at the system level by adding a secondary memory between the DRAM and ASIC as shown in Fig 1b and 2.
- Allows separate fab process optimization for ASIC logic and SRAM memory, which drives increased yields and lower cost.

Due to thermal considerations, ASIC floorplanning is critical. An architecture comprising HSS on the chip edge, many small compute cores in a ring around the 3D SRAM footprint, and other

978-1-5386-6508-4/19/$31.00 ©2019 IEEE

low-power supporting IP blocks directly under the 3D SRAM, as shown in Fig 3, allows for fast memory access while maintaining thermal solutions for heat removal of the ASIC.

Chiplet Packaging

A. Heterogeneous Integration

Standard SoC designs for AI applications include all functional IP blocks placed side-by-side in a monolithic design. An example of an IoT edge computing SoC from Qualcomm is shown in Fig 4 [4]. One potential downside of a monolithic approach is that each functional block may have an optimal performance and cost at different nodes. Sourcing each functional IP block from different nodes as chiplets allows each IP block to be fully optimized for performance and yield. This requires an advanced packaging solution to provide connectivity between these chiplets, such as the concept under development by UCLA, which utilizes sub-10µm pitch interconnects to connect known good chiplets of various technology nodes through a Si interposer fabric [5]. The schematic shown in Fig 5 illustrates the concept. Although the chiplet interconnect formation itself falls well within wafer process capabilities, the chiplet placement accuracy on the Si interposer is difficult at such fine pitch, and is driving significant development by the tool suppliers. Current state of the art is 10µm pitch capability (although not in volume production), driving to 5µm in the future.

B. Scalable Computing

The large investment required for an advanced node logic mask set is driving new market opportunities in scalable computing, where small chiplets can be placed in arrays through advanced packaging to enable multiple products with one design (i.e. data center and edge computing applications). This concept is illustrated in Fig 6. These chiplets can be packaged in an MCM configuration on a laminate, using standard wiring rules between chiplets. The required inter-chip interface is an XSR SerDes type interface to achieve ~2Tbps per mm of chip edge. While the SerDes can achieve impressive bandwidth density, it comes at the cost of power (~1-2pJ/bit) and latency (200-300ns). In order to decrease power and latency, a parallel interface is required. This drives advanced packaging solutions, such as high density fanout (HDFO). The use of a fine pitch (1-2µm L/S) polymer based RDL for chiplet connections can enable the same ~2Tbps bandwidth with a simple parallel interface at much lower power (~0.5pJ/bit) and latency (~6ns). Although a Si interposer based option may allow for slightly higher wiring density, the cost increase over HDFO makes it a less attractive solution.

3D Interface Design Considerations

Enablement of vertical 3D/TSV SRAM connections requires the use of a simple and flexible die-to-die interface, which is needed to place dedicated SRAM blocks and access points for each individual core. This can be satisfied by a highly parallel interface design with multiple I/O, similar to a chiplet interface. I/O and power delivery of the memory dies in a 3D configuration is achieved with TSV connections through the bottom ASIC die. In order to minimize the area lost in the bottom die to the interface, the TSV size and pitch must be as small as possible. Preferred dimensions are 2x20µm TSV with a 7.5µm minimum pitch, maintaining an aspect ratio of 10:1, which is optimum for manufacturing. This results in a final wafer thickness of 20µm per memory slice, which is thin enough to minimize per-pin capacitance loading, but thick enough for manufacturability, and avoids thinning induced device shifts. The area between the TSVs can be utilized for ESD protection or voltage regulation content. It should also be pointed out that stacking higher levels of memory allows amortization of the memory I/O for a given footprint, which is an advantage over placing more memory laterally.

Conclusions

Two advanced packaging solutions have been described to support the next wave of AI products expected to be released. 3D packaging allows for significantly higher processor memory capacity and bandwidth at a lower power penalty. Chiplet packaging enables heterogeneous integration of multiple nodes, as well as scalable options to create several product offerings from one chip design. Although SerDes interfaces are available to support these die-to-die connections, a simple parallel interface is proposed to achieve roughly equivalent bandwidth with significantly lower power and latency, made possible by the short interconnect wiring through the advanced packaging technologies described previously.

References

[1] A. Andrae, "Total Consumer Power Consumption Forecast," https://www.researchgate.net/publication/320225452

[2] V. Sze, et al, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," Proc of IEEE, Vol. 105, Issue 12, p 2295.

[3] L. England & T. Letavic, "Heterogeneous Integration: Architectures and Technologies for Efficient Learning Systems," Stanford System X Workshop, May 2018.

[4] A. Mehta, "On-Device AI Processing for IoT Applications," Linley Fall Conference, Oct 2018.

[5] A. Bajwa, et al, "Heterogeneous Integration at Fine Pitch (≤ 10 µm) using Thermal Compression Bonding," 67th IEEE Electronic Components and Technology Conference, May 2017.

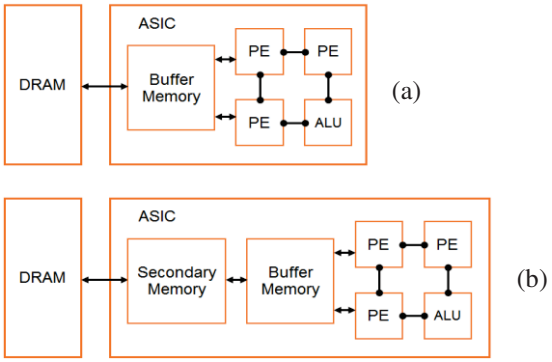

Fig 1: (a) Functional schematic of a typical artificial intelligence system, and (b) addition of secondary memory (i.e. 3D SRAM) to the system. [2]

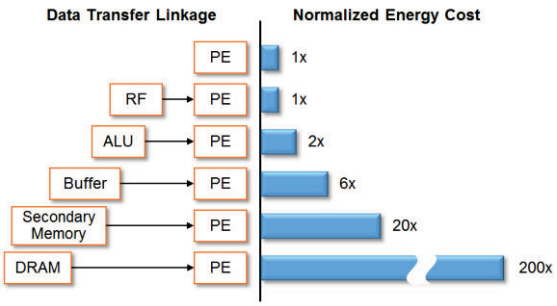

Fig 2: Normalized energy cost to move data between the processing element and other components of the artificial intelligence ASIC. [2,3]

Table 1: Bandwidth per energy for various memory types available for use in AI applications.

Memory Type	Bandwidth per Watt (GBps/W)
GDDR5 (Off Package)	16
HBM2 (2.5D)	28
HBM2 (3D)	28
SRAM (3D)	267

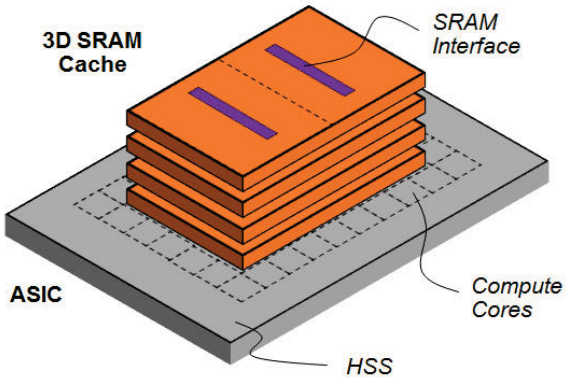

Fig 3: Artificial intelligence architecture containing many small compute cores in a ring configuration around multiple levels of dedicated off-chip SRAM cache.

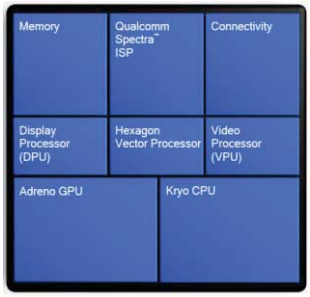

Fig 4: Example monolithic SoC floorplan of an IoT/AI Edge device from Qualcomm. [4]

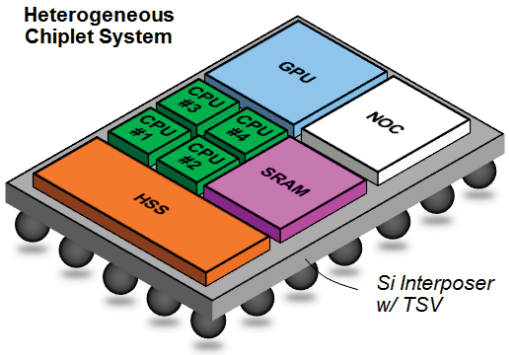

Fig 5: Illustration of heterogeneous chiplet based AI package solution. Each color represents a different fab technology node source.

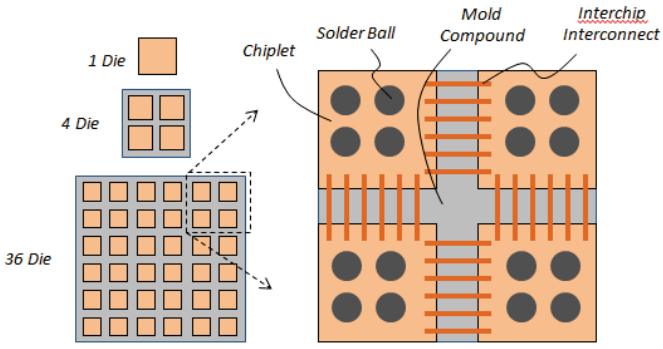

Fig 6: Illustration of scalable computing packaging concept using HDFO packaging. The Inter-chip Interconnect utilizes minimum feature size RDL wiring.