Image Sensors in 3D stacking technology: Retrospective and perspectives from a digital architect point of view.

Jérôme Chossat
STMicroelectronics, Grenoble, France

Abstract

Over the past 10 years, most image sensor volume production has shifted from monolithic towards 3D stacked multi-layer processes. This has been quite a revolution, opening game-changing opportunities in terms of device content and capabilities. In this presentation we will step back and recall the reasons and fundamental choices made at the time embarking in stacked sensor design. The advantages of 3D stacking will be examined, and the opportunities and associated complexity of these sensors will be discussed from a digital architect point of view. Specifically, the management of power and the integration options and capability for on-chip complex image signal processing. We will then have a look at opportunities for integration in image sensors of Artificial Intelligence, trying to scope what is reasonable, what are the limits, and what could be relevant criteria for AI integration in present and future image sensors.

3D stacking rationale

When moving to 3D stacking, some fundamental choices have been made in STMicroelectronics: First, wafer level stacking (in opposition to a die on wafer) has been chosen to cope with large volume production. Second, hybrid bonding has been naturally adopted for die interconnect process, as it is very convenient with BSI process and allows fine connection pitch. Third, top tier has been allocated only to pixels to get optimized image quality with a minimum amount of process steps.

One of the major drivers for 3D stacking, has been to guarantee a certain level of independence between the pixel process and the CMOS logic process. Before 3D stacking, it was required to develop and qualify both processes at the same time, leading to significant trade-offs and dependencies. 3D stacking allowed us to overcome this limitation and paved the way for easier access to advanced CMOS node. CMOS 40nm technology was the first process we used for the bottom tier, which provides a very good balance between analog requirements and digital design (power, speed, and density).

One additional advantage of 3D stacking technology is the capability to get top die limited devices, leading to minimal X,Y size for the imager product. Furthermore, it allows to design optically centered devices, which is not easy with monolithic process and leads to complex U-shape or L-shape digital floorplans, limiting routing efficiency and affecting performances and power consumption. With 3D stacking, the digital shape is usually a rectangle, as depicted in Figure 1, significantly simplifying the physical implementation with better routing efficiency and performances. More recently, a split of the pixel in 2-layers (implementing a part of the pixel transistors in the second layer) emerged for shrinking further the pixel size [1]. This comes with a cost impact: Shrinking the pixel using an additional layer can improve the physical dimension of the device or can improve the pixel performance but is not saving pixel silicon area. Also, the footprint of the array on the bottom die brings similar floorplan shapes than for a monolithic device and inevitably leads to a bottom die limited device, which can then only be avoided with a triple stack solution [2], or with a 3-layer sequential bonding solution [3].

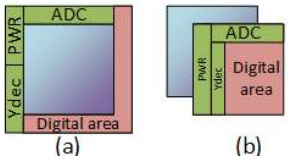

Figure 1 Evolution of floorplan with 3D stacking.
(a) Typical monolithic image sensor floorplan,
(b) Typical 3D stacked 2-layer image sensor floorplan.

Digital node for image sensors

The choice of the CMOS node for a 3D stacked imager is not a process race following Moore's law and must be based on several practical considerations and constraints:

- The device die size is (at minimum) given by the top die, which results from the imager resolution and pixel size, plus a couple of array-to-die-edge constraints.
- The bottom die contains a significant ratio of analog circuitry (Power management, A2D conversion, array control) requiring a process adapted to analog design and is not scaling down along with smaller process geometry (or only marginally).
- Some specific digital constraints can only be achieved with small transistor geometries, such as high-speed interfaces, high memory density, or low digital power consumption.
- The image sensor segment is a highly competitive market, and the process cost is a key parameter, so process choice must not overshoot.

The above considerations are indeed limiting the tendency to shoot for a very dense bottom layer technology. In practice the range of bottom die process is commonly from 40nm to 18nm in production, and down to 14nm in publications [4].

When accessing smaller technology nodes, a natural practice is to use faster clock speed to increase processing throughput. However, in the context of image sensors, for which power efficiency is a key indicator, other strategies have been proved to be relevant: Among them analyzing benefits of architecture parallelism is an interesting path.

The drawbacks of increasing clock speed are two folds: First it tends to increase the ratio of high drive cells in the device leading to a high level of leakage and impacting dynamic consumption through an increase of buffering strength. Second it can lead to very high local power density creating voltage drop issues during physical implementation phase.

The first point is illustrated in Figure 2, showing leakage and dynamic (per Mhz) consumption trends versus synthesis target frequency for an image signal processing IP in 28nm FDSOI process. This is design dependent, but we can see, from this example, that doubling the frequency has led to increase the leakage by a ratio of x13, and the dynamic (per Mhz)

consumption by a ratio x1.4.

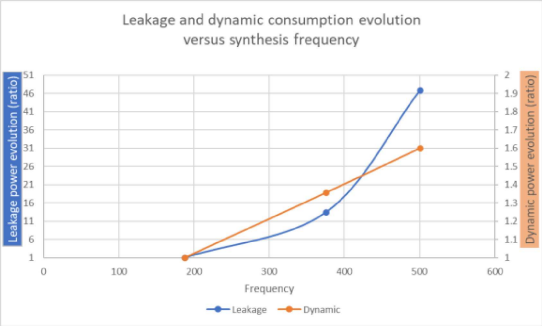

Figure 2 Leakage and dynamic power evolution versus synthesis frequency

Processing parallelism is a way to avoid this, allowing to decrease clock frequency and leading to a significant reduction of the amount of high leakage cells and of the total number of buffers. Of course, parallelizing has a mechanical impact on the quantity of cells, therefore would have tendency to increase the leakage as well, but in much lower proportion (times 2 worst case compared to times 13, in the example of Figure 2).
Listing overall benefits:

- Lowering the timing closure frequency, hence limiting large size buffers insertion, and therefore saving dynamic and leakage consumption.
- Possibility to mutualize intermediate results from the parallel computation (and so to decrease the net number of operations) saving dynamic power and leakage (if the algorithm is designed with this prerequisite).
- Running at a lower speed makes it possible to decrease power supply voltage, hence saving on both dynamic power (by a factor of square the voltage reduction ratio) and on leakage by a factor between x^2 and x^3 of the voltage reduction ratio).

This is an effective manner to turn available area into power reduction with no additional cost on a 3D stacked device.

Bringing added value with on chip processing functions

Still under the hypothesis of top die limitation, there is, in many situations, the possibility of bringing added value to the product by integrating dedicated on-chip processing within the area available on the bottom tier. The following cases must be considered:

- The processing is specific to the data delivered by the sensor, so custom processing is required, and if not done inside the sensor, it would have to be done by an external processor, or through a heavy SW processing in the host.
- The processing can benefit from a close coupling with the imager focal plane to deliver a unique functionality (e.g., specific & efficient readout modes for always-on detectors)
- The processing done on-chip would lead to an overall lower power bill at system level (when aggregating the transmission energy saved and the external processing energy)

I can mention as examples some realizations done in ST products illustrating these cases:

- CFA transposition for supporting RGBIR - integrated in ST product VD1940: Processing the RGB-NIR data to produce a full resolution IR image and a Bayer CFA image.
- Low power motion detection - integrated in ST product VD55G0: Detecting motion and waking-up the device.
- On-chip iToF processing, as described in [5]: Capability to compute the depth map inside the image sensor.
- Optical flow, integrated in ST product VD56G3: Optical flow vectors provided along with the image, making possible to compute camera ego-motion and to analyze intra scene motions with a low-cost microprocessor.
- Always-ON detector [8] developed in collaboration between STMicroelectronics and CEA/Leti and able to detect a face reliably (Accuracy>95%) with a device consumption of 6uW at 5fps.

Complexity with device scalability

Analyzing the area breakdown available in a 3D stacked image sensor bottom die, we can see that the percentage of the area available for digital functions is strongly linked with image sensor resolution. In consequence, over a given product family covering multiple resolutions, the digital content will have to be adapted versus the resolution, and consequently lower resolution may not integrate full features (or higher resolution might be left quite empty). Indeed, as depicted on Figure 3 the percentage of space for the canonical image sensor functions (Array control, conversion, and power management) becomes quickly predominant when lowering resolution, while space available for digital integration decreases quickly.

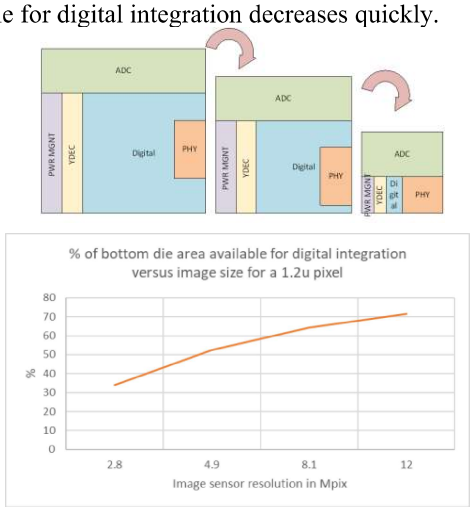

Figure 3 Area available for digital processing versus resolution in a 2-layer top die limited image sensor.

This fact must be balanced with a couple of additional rules defining the minimum distance between pixel array and die edge, which clamps the achievable minimum size, but the trend is however true. The same happens at constant resolution when shrinking the pixel size.
So practically, starting from a given product, designing a cost reduction by moving towards a lower resolution array, or decreasing the pixel pitch, implies to adapt the feature list or to select a higher density digital node (which sounds not so relevant when designing a cost reduction, and will imply redesigning all the analog functions). This is not true if the floorplan of the initial product was quite empty, but in this case,

we can argue the initial product cost optimization.
Obviously, we can imagine breaking the rule for top die limited system, but bottom die expansion over the top die size leads to a wasted area on the top die, as shown in Figure 4.

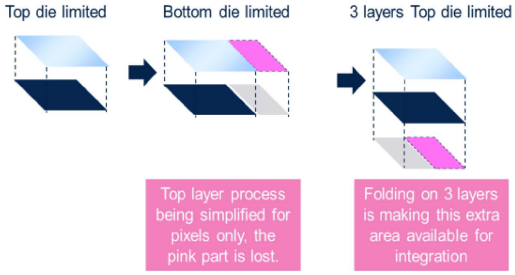

Figure 4 Area wasted on top die brought by bottom die expansion.

Interestingly, it can be shown that in some cases, a 3-layer 3D imager could have led to a lower silicon cost than an equivalent 2-layer with same internal content. This is what is presented in Figure 5. For information, this cost modelling considers the technology cost for each layer and the compound yield.

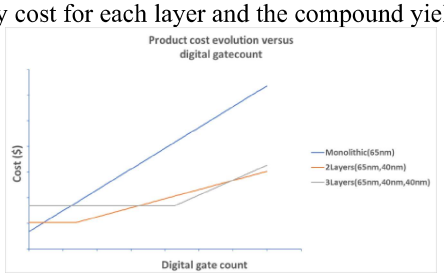

Figure 5 Device cost versus digital gate count in 3D stack device for monolithic, 2-layer and 3-layer device

This problematic is specifically relevant when thinking about integration of Artificial Intelligence in an image sensor. Indeed, this is a case for which resolution may be quite low (as deep neural networks cannot process large input resolutions), while digital content might be quite large. This is what will be analyzed in the next section.

Perspectives for intelligent sensors

Integration of machine learning based computer vision in image sensors has been demonstrated with monolithic sensors, but with significant limitations, e.g., build-in pixel complexity [6] (preventing any future pixel shrink), hardwired features extraction [7], or hardcoded machine learning processing [8]. Such cannot meet the performances of state-of-the-art DNN solutions for analyzing a scene and extracting semantics or detecting specific objects with high accuracy, and do not have the right level of flexibility to adapt to any change of the use case in the field. Over the past 15 years, many network architectures have been proposed. Most of them have been driven by the ImageNet contest [9], which is a complex classification problem. Therefore, a large variety of the proposed networks require high computing performance and are targeting NPU and GPU processing and not on-the-edge image sensor integration. Among those, what is possible to integrate in a 3D stack sensor? An estimation of the available space for digital functions, in a 3D stacked top die limited image sensor is shown in Figure 6 versus image resolution and pixel pitch. One can see that, except in the case of large resolution and large pixels – which seems to have low chance of success in a consumer market – a reasonable digital area available is much below 10 mm2.

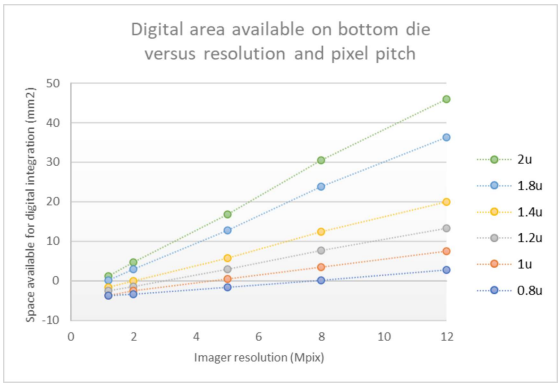

Figure 6 Estimation of available digital area in a 2layer 3D-stacked imager in function of resolution and pixel pitch.

Among this budget, let's consider that a couple of mm2 is required for canonical image sensor digital functions (e.g. descrambling, dark calibration, basic ISP functions, high speed interface, device housekeeping, PLLs,..) then it remains a budget of about 8 mm2 for hosting a potential AI inference solution for the largest sensors, and at most a very tiny area for low resolution and small pixel sensors. If we refer to the representation of neural network performance versus complexity and number of coefficients [10] reported in Figure 7, we can identify a large gap. This publication was done in 2018 but is still quite valid from area and parameter size considerations. The size of the bubbles represents the number of network parameters. Only a very few network models are below 5M parameters - 40Mbits for 8bit per parameter- which represents approximately a budget of 8mm2 by themselves in a 28nm process. Such networks are surrounded with the red dashed line on the graph on Figure 7. Even considering recent work from the TinyML community, most computer vision networks designed for on-the-edge processing are in the 0.5-5M parameters range which represents approximatively 1 to 8 mm^2 in 28nm process. And this is only considering network parameters storage, and not even considering memory requirements for activation values (which by thumb rule are at least 2 to 4 times the number of parameters storage), nor considering area for AI accelerator engine.

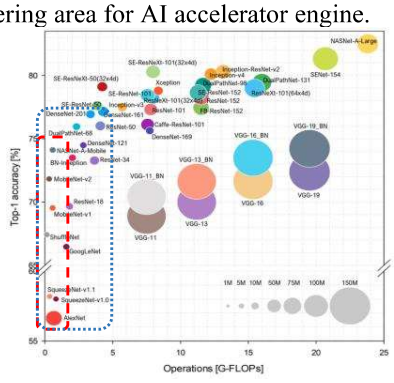

Figure 7 Classification accuracy versus Operation[G-Flops] and parameters [MB] from [10]
Red dashed line are networks <5M parameters.
Blue dotted line area <15M parameters

As reference point, an image sensor embedding a flexible AI

inference engine and containing 9MB internal memory has been published [11]. This was achieved thanks to a quite large resolution (12.3Mpix), a large pixel (1.55um), an advanced technology node (22nm) and at the expense of a significant bottom die limitation: Bottom die is more than 2 times larger than the imager array. Incidentally, we could compute that partitioning this device as 3-layer top die limited, still with a 22nm technology, would have significantly increased the available area for digital integration, while not impacting much the product cost. Reporting this on Figure 7, we can see that this would extend the possible supported networks range to the region surrounded with blue dotted line. In this work, the AI integration cost is high, and it might be difficult to justify the advantage versus using an external processor. Running AI in a sensor must not aim at competing with external processor: Objectives and constraints are different, and it is paramount to identify the rationale and advantages for embedding AI on the sensor. We can propose the below criteria:

- Tangible benefits of a close coupling between the focal plane and the network processing (specific readout modes)
- Capability to obtain a lower power envelop at system level with internal processing (saving in data transport).
- Stringent privacy requirements strictly preventing to send the image out.
- Specific latency requirements.
- Functionality fully embeddable in the limited area available in the bottom die of a 2-layer sensor (at no extra cost)
- Power consumption is maintained in an envelope compatible with image quality constraints.

These are typically criteria having high importance in battery powered always-On systems (power), consumer robotics (latency, privacy and cost), PC presence monitoring (size, privacy, and cost), or AR/VR sensors (size, power, cost).

However, an optimized integration seems only possible through a complete vertical analysis: Rationale for integration, specification of the class of applicative use cases to support (e.g. multi-object detection, classification, segmentation,..), definition of the degree of on-chip flexibility required (max number of parameters, quantization, type of layers supported, network architectures supported,..), but also development of tailored neural network adapted to the target use cases, and specification of adapted hardware acceleration engine to operate the network within target area and power constraints.

Conclusion

3D stacking has opened opportunities for digital integration, improving device power consumption and making possible complex processing and analysis functions within an image sensor. AI solutions are of another complexity scale and bring integration challenge for a 2-layer 3D stacked top die limited device, as the space available on the bottom die depends on pixel pitch and imager resolution. Large image sensors can host small generic AI networks, running in highly subsampled mode. Small image sensors can host only ultra optimized micro-AI networks or must accept extra area cost due to bottom die expansion. Very aggressive digital technology is not helping much, as ratio of analog on the bottom die is usually significant for small imagers. The generalization of 3-layer stacking will facilitate integration, enabling more on-chip memory, but cost impact and flexibility might be questionable versus using an external AI processor. Identifying the benefit for integration in the sensor is key, and we can quote specific cases for which AI integration is bringing advantages and differentiation. However, efforts for vertical optimization are mandatory, from the scoping of application use case through the network architecture definition and down to the design of the hardware acceleration engine.

References

[1] G. Park et al., "A 2.2µm stacked back side illuminated voltage domain global shutter CMOS image sensor," 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA,USA,2019,pp.16.4.1-16.4.4, doi:10.1109/IEDM19573.2019.8993606.

[2] S. -S. Kim et al., "3-Layer Stacked Voltage-Domain Global Shutter CMOS Image Sensor with 1.8µm-Pixel-Pitch," 2022 International Electron Devices Meeting (IEDM), San Francisco, CA,USA,2022,pp.37.5.1-37.5.4,doi: 10.1109/IEDM45625.2022.10019515.

[3] F. Guyader et al., "3-Tier BSI CIS with 3D Sequential & Hybrid Bonding Enabling a1.4um pitch,106dB HDR Flicker Free Pixel," 2022 International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2022, pp. 37.4.1-37.4.4, doi: 10.1109/IEDM45625.2022.10019432.

[4] D. Yu et al., "14nm FinFET process technology platform for over 100M pixel density and ultra low power 3D Stack CMOS Image Sensor," 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2019, pp. 8.1.1-8.1.4, doi: 10.1109/IEDM19573.2019.8993567.

[5] C. Tubert and al. STMicroelectronics IEEE sensors 2022 1.4kDots consumer LiDAR up to 10m based on indirect Time-of-Flight sensor

[6] M. Lefebvre, L. Moreau, R. Dekimpe and D. Bol, "7.7 A 0.2-to-3.6TOPS/W Programmable Convolutional Imager SoC with In-Sensor Current-Domain Ternary-Weighted MAC Operations for Feature Extraction and Region-of-Interest Detection," *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2021, pp. 118-120, doi: 10.1109/ISSCC42613.2021.9365839.

[7] H. Song, S. Oh, J. Salinas, S. -Y. Park and E. Yoon, "A 5.1ms Low-Latency Face Detection Imager with In-Memory Charge-Domain Computing of Machine-Learning Classifiers," 2021 Symposium on VLSI Circuits, Kyoto, Japan, 2021, pp. 1-2, doi: 10.23919/VLSICircuits52068.2021.9492432.

[8] A. Verdant et al., "A 3.0µW@5fps QQVGA Self-Controlled Wake-Up Imager with On-Chip Motion Detection, Auto-Exposure and Object Recognition," 2020 IEEE Symposium on VLSI Circuits, Honolulu, HI, USA, 2020, pp. 1-2, doi: 10.1109/VLSICircuits18222.2020.9162854.

[9] Olga Russakovsky*, Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. (* = equal contribution) ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015

[10] Bianco, Simone & Cadène, Rémi & Celona, Luigi & Napoletano, Paolo. (2018). Benchmark Analysis of Representative Deep Neural Network Architectures. IEEE Access. 6. 64270-64277. 10.1109/ACCESS.2018.2877890.

[11] R. Eki et al., "9.6 A 1/2.3inch 12.3Mpixel with On-Chip 4.97TOPS/W CNN Processor Back-Illuminated Stacked CMOS Image Sensor," 2021 IEEE International Solid- State Circuits Conference (ISSCC), San Francisco, CA, USA, 2021, pp. 154-156, doi: 10.1109/ISSCC42613.2021.9365965.