Chapter 1

Introduction

Present day electronic systems are able to perform a variety of complex functions, attributed to their inner structure consisting of a network of highly specialized sensing, actuating, emitting, storing and computing devices. These are typically interconnected electrically and mechanically over several levels of hierarchy: Chips are packaged to components, then soldered onto a printed wiring board, which usually is the main systems carrier. Over the years, with rising complexity, the number of signal paths per component has increased, and today, performance is no longer exclusively limited by its components, but by the communication paths on and in-between its components. New solutions in interconnect technology are thus required.

1.1 Motivation

The above statements can be deducted from tendencies observed in present day's integrated circuits. To name but a few, the convergence of technologies and rising demands on on-chip wiring are discussed in the following.

1.1.1 Multipurpose Technology Products

Today — especially in mobile systems — market demands a high level of functionality in a small, portable, or even "hand held" application. On the electronic component level, this means, that as many functions as possible need to be integrated into one chip or into one package, since board space is very limited.

The problem that arises is, that such complex systems employ a variety of technologies, i. e. a signal processing unit, a storage unit and an RF unit, which are normally manufactured in a technology specific way using separate dedicated production processes and optimal host wafer substrates. At present, two main system integration principles are in practice.

One way is to extend the IC manufacturing processes in a way that allows a monolithic chip to host areas of different technologies. This is known as system-on-chip integration, SoC, and has the advantage of fast and efficient on-chip interconnects, as long as performance is not limited by long global interconnect lines on large chips (see next section), or by compromising with other technologies. The interconnects between the different parts of the system are on-chip, and do not need to be addressed further. Clearly such systems need to be designed carefully for their many interfering parameters, which define the combined implementation of multiple technologies. The controversy of such practice

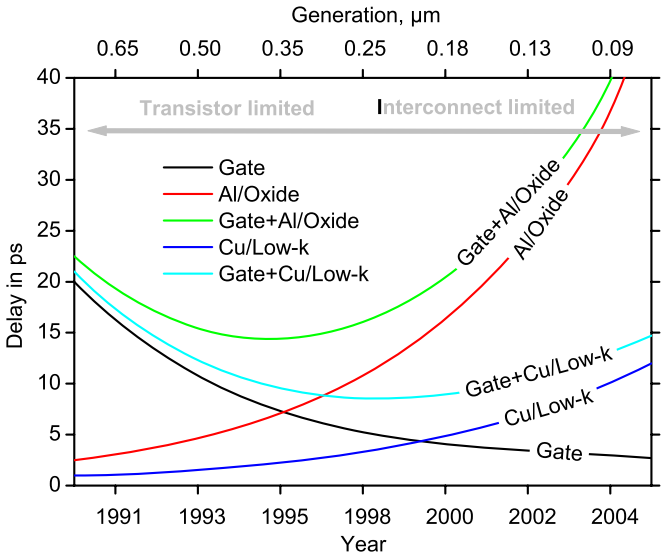


Figure 1.1: On-chip wiring problem: Interconnect delay time dominates over all delay time. Diagramm extracted from Sun et al. [1]

is a high development cost or long time-to-market, and yield sensitivity when increasing the number of process steps.

Another practice is to place chips from different source technologies as close together as possible and combine them into a heterogeneous system. This is generally understood as system-in-package, SiP, which summarizes all possible solutions of interconnecting chips in a hybrid package. This allows to develop and implement the different parts of the system independently or reuse them modularly in a flexible way, resulting in a short time-to-market. Clearly, performance heavily depends on how the chips are placed and interconnected inside the package, and there exists a need to develop efficient methods that result in short chip-to-chip interconnect paths with a low signal delay time.

1.1.2 On-chip wiring demands

In homogeneous chips, the observed tendency of increasing chip areas i. e. for microprocessors and shrinking transistor dimensions have lead to a turning point, where the performance is no longer determined by transistor figures, but by the delay of long (global) interconnect paths.

Figure 1.1 illustrates the projections of Sun et al. [1] from 1997, in which the signal delay time caused by on-chip interconnect parasitics, mainly resistivity and capacitance, will increase in future IC generations, whereas delay time caused by transistor gate parasitics will decrease. The introduction of copper [2, 3] as a replacement for aluminum interconnects and low-κ dielectrics [4] as replacement for SiO_2 has been beneficial to the overall delay time in integrated circuits, but has only postponed the raise of the delay time by a few years (see Figure 1.1). As stated by Tu [5], the RC-delay time is not the only performance limiting factor associated with on-chip interconnects, but also degradation by electromigration in present day copper interconnects.

A potential solution to serve future wiring demands is to shorten the overall intercon-

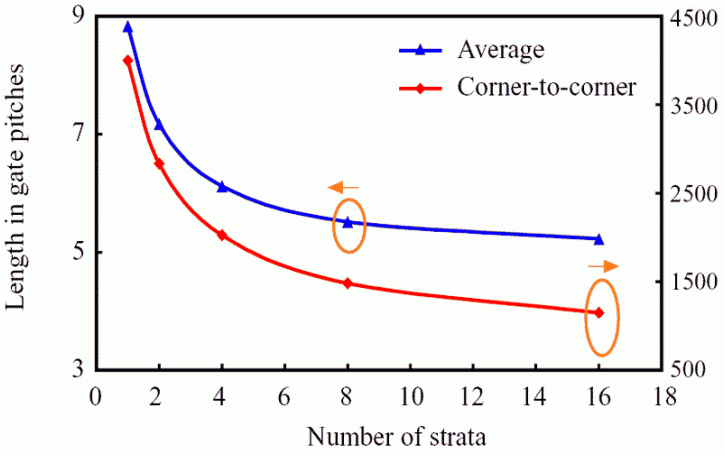


Figure 1.2: Interconnect length for multi layered (multi strata) circuits as modelled by Meindl et al. [6]. The occurrence of long interconnects drops significantly with more than 4 layers

nect length by partitioning a large chip into several smaller chips, which of course then need to be efficiently interconnected off-chip.

1.1.3 Three-dimensional integration

The shortest interconnect path from an arbitrary point on a chip to one on another chip can be achieved by the stacking of chips, thus using the third dimension for interconnecting.

Studies on stochastic models of different scenarios of multilevel (1 to 16 strata; stratum = level, layer) circuits with direct interlevel connection paths [6] lead to the conclusion, that the length of interconnects in only 4 layers of such 3D integrated circuits is already considerably shorter as compared to non 3D architectures. Shown in Figure 1.2 are the results obtained by Meindl et al. [6] for a model of a random logic network with ca. 4 million gates distributed and interconnected over 1 to 16 levels. The curves for average (mapped to the left axis) and corner-to-corner (right axis) interconnects show a similar exponential behaviour. The curves also imply, that 3D-integrating of more than 8 layers will have no substantial benefit on shortening interconnect lengths in random logic networks.

The random logic network model can be considered rather pessimistic for it does not represent a 3D-optimized architecture or does not take into consideration structures like buses and cell arrays which would probably benefit most from 3D-integration for their regular 1D and 2D patterns. Systems, where 3D-integration is considered most beneficial, are thus memory applications and image processing applications.

3D-integration is an emerging technology that has the potential to solve the on-chip interconnect problem by partitioning a large chip into smaller ones, which can then be interconnected vertically using the shortest path. It also has the potential to satisfy technology convergence needs by allowing heterogeneous integration of chips from different sources.

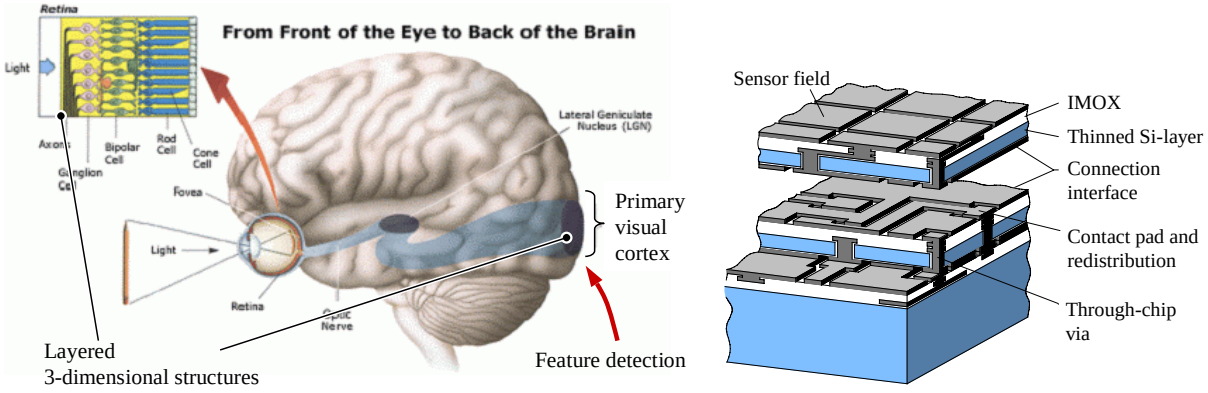


Figure 1.3: Biological vision system and its possible hardware replication as a stack of silicon chips bearing a neural network. Graphics from [9]

1.1.4 Biologically inspired neural processing

Today's need for 3D-systems mainly originates from two fields of application, memory storage devices and image processing devices. Emerging technologies, such as neural network computing devices have an even higher need for 3D-integration technologies, because as many nearest neighbour neural cells as possible need to be interconnected in order to function efficiently. Pulsed neural networks, as implemented and investigated at Infineon Corporation and Dresden University of Technology [7, 8] operate at a "low" frequency in the kHz range, where clock or signal interference and heat dissipation need not be considered.

The architecture of such a neural imaging system is biologically inspired (see Figure 1.3), and can basically be described as a meshed cube with knots that are in charge of specific functions, as to in which layer they are located. As a replication of the biologic visual system these layers can have the function of the retina (first layer: image sensor), ganglion cells (second and other layers: feature extraction) and cortex (last layers: image processing and storage). Such an architecture is easy to implement into hardware because it is straight forward and highly parallel with repetitive patterns (cells).

Because of its low operation frequency, neural processing is an evolving strategy for low power efficient computing in the domain of signal preprocessing. Most arithmetics in preprocessing and filtering of images consist of convolutional operations [10], for which a neural network with its inter and cross connections is an efficient architectural configuration. Research, related to the analog VLSI implementation of neural cells in an outer context of this work, has been conducted at the University of Dresden [11], supported by simulations and coordinated from Infineon Corporate Research.

From an application point of view, motivation of work related to this thesis is thus to provide a technological platform for the implementation of a 3-dimensional neural network, which replicates a biological vision system. From a technological point of view, motivation is given from a converging technologies background, where a sensor technology

a) b) c)

Figure 1.4: Examples of peripheral 3D-interconnect technologies: a) Peripheral wire bonding by STMicroelectronics [12], b) NEO stacking by Irvine Sensors Corporation [13], and c) Chip-in-polymer by Fraunhofer IZM Berlin [14]

needs to be combined with a signal processing technology, no matter what technology is host to the system's components.

1.2 Interconnect needs for 3D-integration

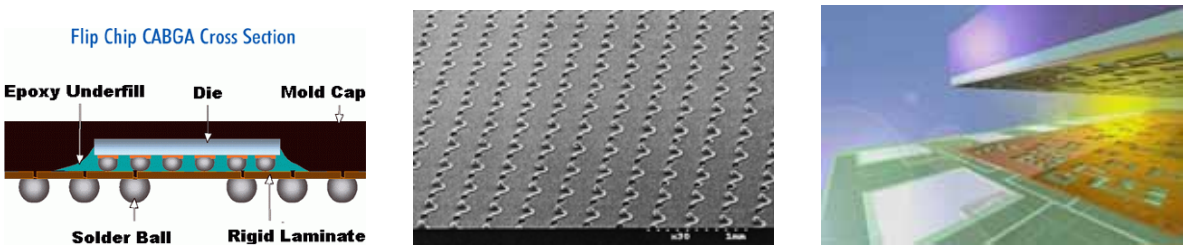
The term interconnect is generally used for defining an electrical connection between two functional entities within a system. It mostly refers to on-chip electrical connections or to chip-to-substrate connection. In evolving 3D-integrated systems, interconnects will also be formed between several levels of stacked chips, thus resembling 3D-interconnects.

3D-interconnects can be classified in various ways. The classification used in this thesis, will be whether interconnects are made within the area of the chips (area-interconnects) or whether the chips are interconnected only at the periphery (peripheral interconnects). A further instance of 3D-interconnects can be made up by interconnects that comprise through-chip vias and inter-chip joints.

1.2.1 Peripheral Interconnects

Classically, chips have been connected to other chips or to a substrate at the chip's periphery, i. e. with bond wires or tape (TAB, tape automated bonding). Recent innovations in 3D-integration have made it possible to produce peripheral wire bond interconnects on thinned and stacked chips, as for example presented by STMicroelectronics [12]. The benefit of such practise is sufficient for many applications, however, in the worst case, a point to point connection from the middle of one chip to the middle of another chip still needs to be as long as a chips horizontal dimension.

There also is an unfavourable linking between the serial formation of wire bonds and cost per chip-stack, when a large number of connections needs to be made. Other techniques thus have been developed, where the peripheral connections are made in a parallel process, for example with metal patterns on the sidewall, as done by Irvine



a) b) c)

Figure 1.5: Examples of area interconnect technologies in order of increasing density: a) Flip-chip package from Amkor [15], b) GeorgiaTech's Sea-of-leads [16], and c) Infineon's SOLID face-to-face technology [17]

Sensors Corporation [13] or by embedding thinned chips into polymers with via feed-throughs, as available i. e. from Fraunhofer IZM in Berlin [14]. Demonstrators of the mentioned peripheral 3D-interconnect principles are depicted in Figure 1.4.

Periphery interconnects may make efficient use of the whole chip or package area by means of contact redistribution, but still result in long interconnect paths and a large package footprint. More crucial however, is the limitation they impose on interconnect density, because then, the one-dimensional periphery turns out to be the bottleneck.

1.2.2 Area interconnects

A higher number of interconnects per chip-area can be achieved, when the whole chip area rather than only the periphery is used for interconnecting. In state of the art packaging, flip-chip techniques are employed, that is, diced chips are bonded face-down onto the package substrate (i. e. a rigid laminate) in order to create chip-to-package interconnects (for example flip-chip-package from Amkor [15], see Figure 1.5a).

In more advanced approaches, interconnect density can be increased by depositing the interconnect metallization system onto the chip surface, when still on the wafer, by means of a batch process. At Georgia Institute of Technology (GeorgiaTech), the sea-of-leads wafer-level packaging (WLP) technology has been developed [16], where metal leads are deposited onto a polymer, the combination of which functions as a mechanical spring for stress relief between chip and substrate.

Infineon's SOLID face-to-face technology is targeted to stack two chips rather than interconnecting a chip to a substrate. This allows even higher interconnect densities, because thermal mismatch is much of a smaller problem between to similar chips, and protection and sealing is taken care of at the package level. Examples of the described area interconnect technologies are shown in Figure 1.5. Infineon's SOLID technology has a particular significance to this theses and will be treated thoroughly in chapter 4 on page 55.

What has been described in this section are high-density area interconnects for con-

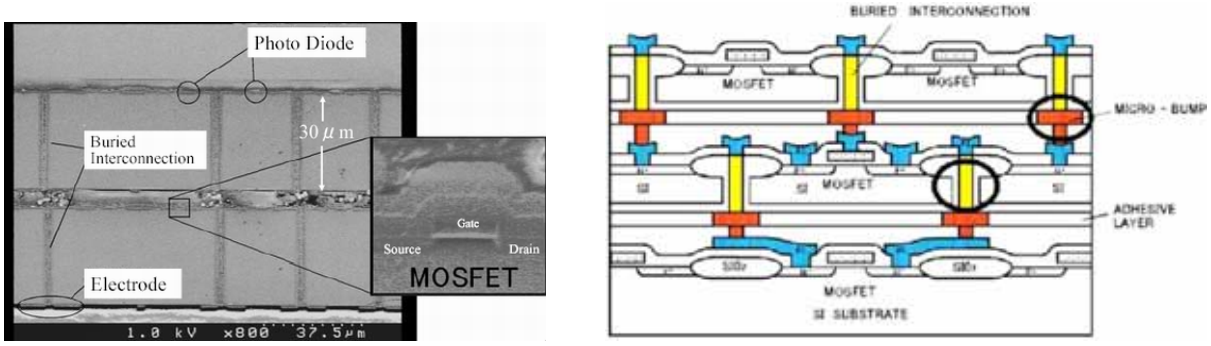


Figure 1.6: 3D integration at Tohoku University, Japan [18]. Small diameter vias with a high aspect ratio were realized using Si deep trench etching and W deposition. Si substrate can then be as thick as 30 µm

ventional 2D-integration, and a stacking technology, with a maximum number of two directly connected chips. For true 3D-integrated systems, with high-density area interconnections between three and more chips, through-chip vias on a micro-scale need to be available. Providing these micro vias in combination with inter-chip joints of similar size has been a subject of research for ca. two decades, and will also be the focus of this work.

1.2.3 Through-chip vias

Through-chip vias provide a connection path from an arbitrary point on the front side of a chip through the bulk silicon of that chip to its backside. These vias, when having a diameter of a few micrometers only, allow the highest density of 3D-interconnects. Methods that provide through-chip vias and allow stacking of three and more chips extend to a rather complex scheme of processing, which mostly includes wafer thinning, DRIE (deep reactive ion etching), via filling, and bonding to a stack of wafers or chips. Laser drilling also can be used to form micro-scale vias, but will not be discussed here for its serial processing nature. Presently known methods of 3D-integration that involve micro scale through-chip vias are introduced shortly in the following.

Tohoku University As demonstrated by the wafer-to-wafer 3D-technology platform developed at Tohoku University, through-silicon vias can be realized with an aspect-ratio (depth to width ratio) of 12:1 (30 by 2.5 µm). The vias were etched 50 µm deep by means of Si deep trench etching and filled with poly-Si and W. First the vias are formed, then the wafer is thinned to 30 µm and bonded to the stack. Thus, when high aspect-ratio via technology is available, the substrate needs only be thinned to 30 – 50 µ, and the thinned wafers need less mechanical support during processing.

The method has been used to fabricate a computing unit with 3D integrated memory [19], an image processing unit and an electronic replication of a biological vision system [18]. Figure 1.6 depicts the vision system integrated into 3 layers of Si chips. The