1 Vertical Integration of High-Performance Processor-Memory Stacks: Motivation & Conception

Since more than four decades Moore's Law dictates the pace of economic transistor integration in the integrated circuit (IC) industry. It postulates the doubling of switching elements every 18 month on a single microprocessor die (chip) [1]. In the same time period several key technologists predicted the end of this trend within one or two generations - this never happened. Down to 130 nm transistor node dimensions integration was manly achieved through "classic" Dennard scaling. Gate-oxide thickness, transistor length and width are scaled with a constant factor improving integration density and clock-frequency at constant power density. Shrinking the switch dimension into the deep-sub-micron range resulted in degraded device performance in all subsequent generations [2]. Hence, the introduction of enhancers such as strained Silicon, high-k metal-gates, low-k wire dielectrics and improved thermal management were responsible to sustain Moor's Law to 32 nm node dimensions in 2010. For future nodes, the need of innovations is becoming more accentuated as indicated by the International Technology Roadmap for Semiconductors[1] (ITRS). They identified vertical integration as a key technology to keep delivering "more for less" also beyond the complementary metal-oxide-semiconductor (CMOS) era.

1.1 Driving Forces and Enabling Technologies

Die-on-die and package-on-package stacking is used in mobile applications to deliver substantial computing performance in a minimal form factor for years already [3]. Individual stratas are fabricated in standard CMOS technology with subsequent wafer thinning before stacking. Unfortunately, the silicon substrate acts as a natural barrier for direct communication between transistors on different levels along the shortest possible distance. Consequently, signals and power have to be routed along the die periphery. This results in severe bandwidth constraints which are not acceptable for high-performance microprocessor applications. They demand for thousands of electrical in- and outputs (IO) as described by Rent's rule for complex systems [4]. Only a method accepting true area-array-electrical interconnects between dies would fulfill the communication needs within a 3D processor memory stack.

The advent of economic fabrication methods to form anisotropic patterns into silicon with subsequent conformal metal filling were the prerequisites to enable the implementation of die stacks using vertical through-silicon via (TSV) communication. Such processes were introduced in the early 90ties and became state-of-the-art in the industry. Deep reactive-ion etching (DRIE), the so called "Bosch process", was developed to fabricate high-aspect-ratio microelectromechanical systems (MEMS) such as sensors and fluidic devices [5, 6]. The IC-industry was challenged to reduce wiring resistance and capacitance by replacing sputtered low-aspect-ratio aluminum wiring with high-aspect-ratio electroplated copper utilizing the dual-damascene process [7] in the same time-period. The avenue for vertical integration was foreseeable with these technologies maturing in a high-volume production environment .

1.1.1 Performance Benefits of Vertical Area Array Electrical Interconnects

Vertical integration opens several new possibilities for the design of microprocessors. The benefits are manifold, such as (i) reduced interconnect wiring length, (ii) improved memory to core bandwidth, as

[1] http://www.itrs.net/

well as (iii) heterogeneous integration with its implications on improved computing performance at lower power and cost [8, 9].

In "classical" transistor scaling, switching delay in transistors is improved as the device dimensions are reduced. In contrast, signal propagation between transistors in the wiring layers is altered [10]. The RC_{delay} scales inverse proportional to the square of wiring pitch P and thickness T, due to increased parasitic capacity and resistance as illustrated in Equation 1.1:

$$RC_{delay} = 2\rho\epsilon(\frac{4L^2}{P^2} + \frac{L^2}{T^2}), \tag{1.1}$$

with metal resistivity ρ, permittivity of the dielectric ϵ, at a wiring length L. The result is an expected 2.5 ps transistor delay compared to a 250 ps RC_{delay} per mm in 32 nm technology [11]. At the same time, the die foot print of high-end server microprocessors is increasing, accentuating the so called "wiring crisis" even further. Additional signal repeaters are needed to assist global wire signal transport affecting total power consumption. So far, evolutionary technological improvements such as low-k dielectrics and an increasing number of metal levels to provide global wires with larger cross-section were introduced. 3D integrated architectures would change the perspective completely by minimizing global wiring length proportional to $1/\sqrt{n}$, with n representing the number of stacked dies [12]. This enhances power efficiency, as well as latency.

A further advantage of 3D integration reduced core-to-cache bandwidth bottleneck especially present for multi-core microprocessors. Their efficiency relays on the low latency to fetch data. Therefore, memory hierarchies with fast but small capacity level zero (L0) cache at closest proximity to the core, followed by more distant and larger capacity L1 and L2 cache are implemented on-chip. An imbalance of cache-to-core ratio leads to core stalling due to "cache-miss" events. This hurts especially if cores have to wait for off-chip data to be fetched, causing hundreds of unused cycles. The unbalanced growth of processor-performance of 60 % per annum, compared to a 10 % per annum memory access-time improvement will continue in case of 2D processor architecture and is the main factor comprising the "memory wall". A heuristic observation postulates a reciprocal dependency of the miss-rate relative to the square-root of the cache size. Considering this, doubling the number of cores at a constant off-chip memory bandwidth results in an 8-fold cache size demand. This leads to ever increasing die footprints which soon will hit the limit defined by the maximal lithographic projection area of about 6 cm^2 [13]. Stacking of cache on cores with TSV vertical interconnects results in massive bandwidth and minimal latency to an enormous amount of memory. Separation of functionality to different layers (heterogeneous packages) reduces also the fabrication complexity of individual dies and therefore allows the implementation of e.g. dynamic random-access memory (DRAM) for L2 cache with 8-fold denser packing density, than static random-access memory (SRAM) cells. Production cost will drop and enables additional functionalities in a chip stack as proposed by the "MORE THAN MOORE" initiative.

Several studies tried to quantify the discussed advantages for high-performance processors. BLACK et al. demonstrated the circuit-level potential of the INTEL CORE DUO™processor with shared L2 cache on top of the cores. The number of cycles per memory access was reduced up to 55 %. Further, a logic-on-logic stack was considered with floating point unit on top of a x86 core. It results in a simultaneous 15 % performance increase and a 15 % power reduction for a SPEC2000 benchmark [14]. KGIL et al. proposed a memory hierarchy change for high-throughput multi-thread applications such as tier 1 web servers [15]. Large capacity DRAM layers are attached at the backside of the logic die with high-bandwidth connectivity replacing on-chip L2 cache. The gained space on the logic die can be used for additional processing cores improving computation throughput while each core can run at modest frequency and therefore improving energy-efficiency. The study concludes with a 14 % performance improvement at a 55 % power reduction over a baseline 2D approach.

Architectural and physical design tools have to be developed to take full advantage of the additional degree-of-freedom from vertical integration. Especially algorithms to optimize TSV arrangement with minimal impact on wiring congestion have to be considered, to take full advantage of global wire reduction. Early studies investigating physical-via-impact conclude with an optimal partitioning granularity on unit level (a unit being a large logical entity such as a floating-point unit) and beyond (Figure 1.1) [16].

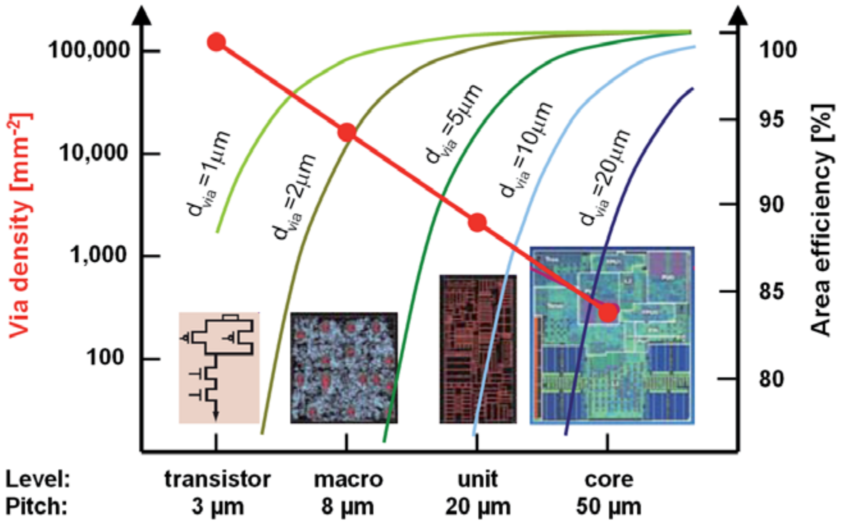

Figure 1.1: 3D partitioning optimum when considering TSV-impact: The available silicon area for transistors (area efficiency) is plotted for different TSV diameters (blue to green lines). Corresponding via densities needed for the communication at different levels is depicted with a logarithmic scale (red line).

This results in TSV pitches ranging from 4 µm to 200 µm^2.

3D integration enables also new business strategies. Intellectual property (IP) sharing is simplified compared to system-on-a-chip (SoC) integration. Functionalities are partitioned to individual stratas and make the designed and fabrication by individual legal entities possible. This approach also enables the use of generic functionality across a product portfolio. Products can be improved and upgraded with minimal cost and time-to-market,simply by replacing one of the layers in the stack, while keeping all the other strata the same. Both aspects increase the number of dies processed for a certain functionality. Therefore, costs are expected to drop according to the principle of economy-of-scale.

1.1.2 Enabling Technologies: Vertical Interconnects - Substrate Thinning - Alignment & Bonding

Two basic approaches were proposed to fabricate a multitude of electric active layers on-top of each other including vertical electrical interconnects. In the *bottom-up* approach a monolithic integration on a single substrate is intended. The bottom-most layer is fabricated using standard CMOS fabrication technology. Subsequent silicon layers are deposited, transistors are defined and are connected by metal wiring layers. It is challenging to achieve high quality subsequent semiconductor layers due to temperature limitations of around 400 °C imposed by the metal wires in place. Epitaxial film quality has to be achieved for high-performance applications to derive charge-carrier-mobilities comparable to bulk silicon. Therefore, localized or low-temperature processes such as laser assisted solid-phase recrystallization [17], metal-induced-lateral overgrowth [18], as well as the implementation of seed agents such as germanium or nickel [19] were proposed to transform the amorphous silicon layers into monocrystalline layers. Another concern is the increasing number of process steps which scales with the number of active layers and comprises device yield.

Hence, the *top-down* approach utilizing 2D die stacking seams to be economically favorable and was proposed in the 80ies due to the aforementioned implications [20]. Each active layer is built on high-quality silicon substrates resulting in optimal transistor performance. Individual layers can be verified prior to stacking, rendering Known-Good-Die philosophy applicable in case of die-to-die or die-to-wafer bonding. Wafer-to-wafer bonding is economical favorable for high-yield devices, due to increased through-put by utilizing the full advantage of batch processing [21]. The main process steps to form 3D chip stacks are: (i) implementation of vertical interconnects, (ii) wafer thinning, and (iii) alignment with subsequent bonding.

[2] This finding will strongly influence the performance of interlayer cooling, as will be discussed in section 1.2.

Vertical Interconnects

Vertical-electrical-interconnects, so called through-silicon vias (TSVs), are the main ingredients in vertical integration and were proposed by ANTHONY in 1981 [22]. TSV fabrication includes the following process steps [23, 24]:

- VIA-ETCHING: DRIE to form high-aspect-ratio holes into the silicon substrate.
- VIA-LINING: Deposition of interconnect supporting layer system - electrical passivation (SiO_2, Si_3N_4), diffusion barrier (TaN / Ta), adhesion (Ti) and seed (Cu) functionality.
- VIA-FILLING: Chemical vapor deposition (CVD) of tungsten (W) or copper (Cu) or electroplating of Cu to form the electrical conductor body.

Four main TSV types are differentiated depending on their implementation during IC fabrication. (i) *Via-first* refers to TSV implementation prior to front-end-of-line[3] (FEOL) processing. Only poly-silicon as a conductor material is compatible with subsequent FEOL processing at temperatures up to 1000 °C. These vias suffer from poor electrical conductivity and cause therefore high resistive losses. TSVs fabricated after transistors implementation, between FEOL and back-end-of-line[4] (BEOL) processing are named (ii) *via-middle*. Due to the relaxed process temperatures during BEOL deposition of $\leq 450\,°C$ conductor materials with improved electrical conductivity such as tungsten (W) and copper (Cu) are applicable. (iii) *Via-last* and (iv) *via-after-bonding* are fabricated typically from copper after completion of the main IC-fabrication.

Feasible TSV heights, diameters, and pitches are strongly dependent on the processes and material involved. They also influence the detailed geometry of the TSV. Economic and void free cylindrical Cu vias can be electroplated with aspect-ratios up to 6:1, using levelers to optimize conformal filling. High mechanical stress is built-up at the TSV-Si interface induced at temperature peaks during fabrication due to the large miss-match in coefficient of thermal expansion (CTE) between Cu (17 ppm/K) and Si (3 ppm/K). The stress increases with the TSV dimension and causes fracture above a critical via dimension. This defines an upper bound for the diameter of cylindrical Cu vias of about 20 µm. For larger diameters an annular via design with a trench width smaller than the critical dimension is proposed. An impedance of 4.4 mΩ was measured at a 50 µm diameter and 150 µm via height with a 4 µm annular ring width [25, 26]. For high-frequency signal transmission the effective impedance of annular and cylindrical TSVs are equivalent considering skin-effects. CVD-tungsten with a low CTE of 5 ppm/K further mitigates mechanical stress, but suffers from reduced electrical conductivity of 9 µΩ cm compared to 2 µΩ cm for copper. Only layer thicknesses below 4 µm are feasible due to the slow CVD deposition-rate resulting in cylindrical vias with diameters $\leq 8\,\mu m$. Trenches with 4 µm width and a depth of up to 150 µm can be filled void free thanks to the high conformality of the process. Large vias will be represented by an assembly of several bar-shape or concentric trenches. A seven bar via with 80 µm diameter and 90 µm height results in an impedance of 25 mΩ [27]. Figure 1.2 concludes the TSV dimensional space for the different technologies discussed above.

Via-first and middle have to be processed by integrated device manufacturers (IDMs) or foundries and cause minimal wiring congestion. Via-last and after-bonding can also be performed by outsourced assembly and test (OSATs) organizations. They are fabricated at lower aspect ratios and cause additional wire-blockage in the BEOL wiring layers. Via-last and via-after-bonding are already available on the market according to a market study performed by YOLE DÉVELOPPEMENT [28]. Via-middle processes are still in the development phase. Via-first are in production for sensor applications, but are not considered as a candidate in high-performance applications due to the low electrical conductivity.

Wafer-Thinning

Wafer-thinning in the range of 10 to 150 µm is necessary to minimize vertical communication distance, as well as to maximize TSV densities. The initial thickness of a 300 mm substrate is 780 µm. First, a grinding step is used to thin the wafer down to about 150 µm. It is followed by chemical-mechanical polishing (CMP) and a plasma etch step, to minimize the risk for defects in the junctions. A desired

[3] The term FEOL refers to all process steps preceding the first wiring metal level deposition, mainly being the formation of the transistors.

[4] The term BEOL refers to all processes concerned with metal wiring fabrication. They follow FEOL processing.

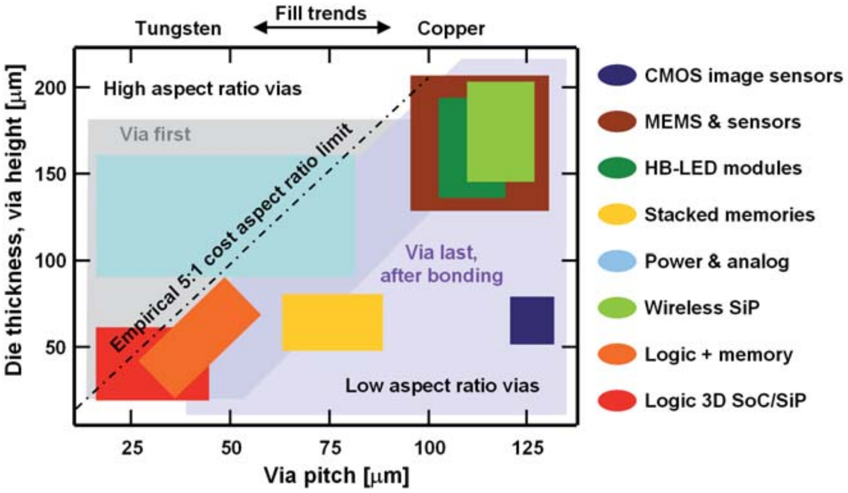

Figure 1.2: TSV pitch and height window for different interconnect density demand, defined by individual applicaitons. Prefered material types and aspect ratios of via first and last are indicated as well [28].

thicknesses of 30 to 150 µm is targeted for bulk devices. Selective wet-etching is used to remove the silicon at the backside of the buried-oxide entirely for silicon-on-insulator (SOI) stratas [29]. The wafer is temporarily bonded onto a handling substrate for mechanical support prior to the thinning process. A spin-on thermoplastic or thermoset compatible with grinding, exposure to chemicals, as well as the stack bonding temperatures performs the temporary bond. Debonding of the handling substrate is done at elevated temperature, where the adhesive becomes viscous and the parts can be separated by a sliding procedure at 200 to 300 °C or by thermal decomposition of the adhesive at 350 to 400 °C [30, 31]. It was also proposed to use electrostatic chucks instead of adhesive bonding, to minimize the risk of thin wafer damage during the debonding step [32].

Alignment & Bonding

Stack-formation can be pursued in serial mode, such as die-to-die (D2D) and die-to-wafer (D2W) bonding or in batch-mode, namely wafer-to-wafer (W2W) bonding. The later method results in highest throughput, but implies a uniform wafer and die size across the chip stack. It is only economical for high yield devices, since stack yield is equal to the die yield to the power of strata numbers. Optimal die size and maximal stack yield on the other hand can be achieved by selecting known-good-dies (KGDs) followed with individual die alignment and bonding as in case of D2D or D2W processing. KGDs can also be reassembled on a carrier substrate, to form a KGD-wafer allowing for high throughput W2W bonding with a resulting high yield.

Face-to-face (F2F) or face-to-back (F2B) strata arrangements are possible[5]. In the F2F arrangement no handling wafers are needed since wafer thinning can be performed after bonding. Furthermore, a high die-to-die interconnect-density can be realized, independent on TSV pitch. At least one F2B bond results for stacks comprised of more than two dies, with the need of TSV and handling wafer.

The main limiting factor for the vertical interconnect density are alignment limitations in the order of ±0.5 and 2 µm for D2W and W2W bonding respectively. Distortions and wafer bowing caused by non-uniform intrinsic stress and non-isothermal temperature distribution during bonding as the root-cause. Tight alignment tolerance for W2W bonding is more demanding, since alignment is performed across a 200 or 300 mm substrate. This tolerance issues result in a practical minimal TSV pitch of 5 µm [33].

The bond itself serves multiple-purpose, such as mechanical integrity, electrical and thermal connectivity at low impedance. Four main bonding systems and combinations are proposed:

Metal-metal thermocompression bonds can be performed between extremely flat Cu-Cu surfaces. Microscopic contact points start to deform at elevated temperatures of 350 to 400°C under the applied pressure. The resulting inter-diffusion of copper across the interface in the subsequent 60 min annealing step under inert nitrogen atmosphere forms a low electrical impedance bond. Mechanical

[5] The active side of the strata carrying the transistors is called face.

integrity is achieved with a copper coverage > 13 %. Additional dummy copper area have to be added at a low number of interconnects, to improve mechanical strength, as well as thermal transfer between stratas [34].

Fusion bonding between dielectric layers such as SiO_2 are typically performed at high temperatures of 1000 °C. Initial Van-der-Waals bonds from pre-bonding are replaced by covalent bonds desired for high bond-strength. Pre-surface treatments using plasma activation or hydroxyl surface termination by wet-chemical processes reduces the annealing temperature substantially to 400 °C. A proprietary process described by Ziptronix using ammonia termination does not rely on heating at all. Covalent bonds are formed at room temperature without chemical byproducts such as water [35]. High surface quality is a prerequisite, achievable through careful CMP. The electrical contact is formed later by via-after-bonding processes or by hybrid bonding as in case of dielectric-dielectric bonds.

Thin film eutectic soldering takes advantage of the liquid metal phase existent during reflow, adjusting for roughness and non-planarity between the bond pads, therefore relaxing the surface quality requirements. Typical pressures of around 5 bar are applied. The most widely reported system is CuSn, discretely deposited using electroforming of tin on top of copper. Copper dissolves into the liquid tin and forms the eutectic phase at reflow temperatures of 260 °C. An increasing reflow dwell time results in solidification of the initially liquid phase. It is initiated by the increasing copper concentration resulting in a high-melting temperature intermetallic compound. This irreversible process allows the formation of bonds at low reflow temperatures, but results in bondlines with higher thermal stability. This enables assembly hierarchies required for sequential die stacking [36, 37, 38].

Adhesive bonding relays on spin-on thermoplastics such as polyimide or benzocyclobutenes (BCB). They become viscous at elevated temperature and form compliant bonds at around 350 °C and applied pressures of 5 bar. A major concern is the poor thermal conductivity of such bondlines of about $0.2\,\frac{\mathrm{W}}{\mathrm{m\,K}}$ rendering a high thermal barrier for heat conduction [29].

Hybrid bonding combines the advantages of metal-metal thermocompression and adhesive bonding. Openings on the metallic receptacles are formed by patterned intermediate adhesive and act as a self-alignment structures. Metallic protrusions from the counter-part are locked into these structures. The polymer film thickness is reduced and the metal bond is performed during thermocompression bonding [29]. “Direct bond interface” is an extension of the pure room temperature oxide fusion bond proposed by Ziptronix. The interface consists of SiO_2 and copper pads in nearly coplanar arrangement. The pads are some nanometers recessed. Therefore a room temperature oxide bond can be achieved as described before. A compressive stress can be achieved on the metal pads due to an increased metal thickness expansion compared to the dielectric layer at elevated temperatures. This results in equivalent metal-to-metal bond quality as derived from thermocompression bonding [35].

To fulfill the requirements of the targeted product chip stack a stringent process sequence has to be defined from the broad variety of technologies available. Some processes will become an industry standard for given applications as technologies mature.

1.1.3 Vertical Integration Product Roadmap

Vertical integration will also have a significant impact on the supply-chain of semiconductor packages. Several players are capable to ramp-up high-volume fabrication of 3D specific technology. Foundries have the highest flexibility in TSV fabrication. Substrate manufacturers and OSATs are only capable to implement via-first or via-last approaches, respectively. Extensive alignment and bonding knowledge is already available at MEMS manufacturers and OSATs. For foundries, these processes are completely new. IDMs have the unique advantage to optimize all process steps during stack formation and do not rely on standardization of components. Whereas the major vertical integration break-through relys on

APPLICATION	WAFER STARTS	WAFER SIZE
CIS	550k	8″, 12″
MEMS / Sensors	158k	6″, 8″
Power / RF	22k	6″
Memory	10k	12″
HB-LED	2.5k	8″

Table 1.1: 3D TSV market share of different applications by wafer starts in 2009 according to YOLE DÉVELOPPEMENT [39].

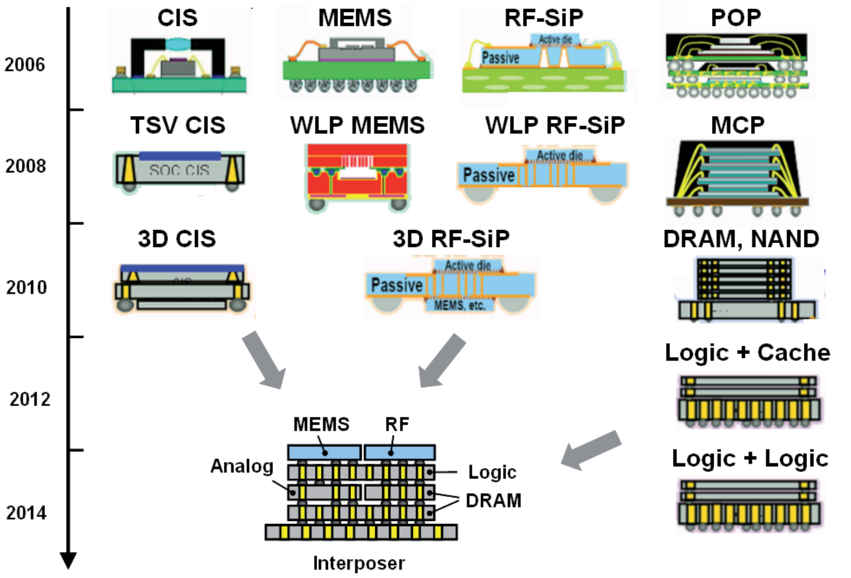

Figure 1.3: Vertical integration roadmap of diverging market segments such as CMOS image sensors, microelectromechanical systems, transceiver modules and memory / logic applications[39].

standardization of IC interfaces and 3D technology to allow stacking of dies designed and fabricated at different suppliers.

CMOS image sensors (CIS) were dominating the 3D TSV market in 2009, taking a share of nearly 3/4 by the number of wafer shipped according to a market study performed by YOLE DÉVELOPPEMENT [39]. MEMS sensor applications accounted for the other quarter. Power, radio frequency (RF), and memory applications, as well as high-brightness light-emitting diodes (HB-LEDs) with integrated TSV technology were brought to the market recently (Table 1.1).

ELPIDA announced to ship first vertical integrated 8-layer DRAM stacks in 2010. The next evolutionary step will be to implement such memory stacks side-by-side with a high-performance microprocessor on a silicon carrier, allowing for high density signal lines between the components [40]. Such system-in-package (SiP) assemblies will dominate the 3D TSV market from 2014 on according to market forecasts. A moderate 10 % share for memory-on-logic and logic-on-logic is predicted until 2015 (Figure 1.3).

About eighteen 300mm foundries are transforming to become TSV compatible world wide. This sites are equally distributed between North America, Europe, South East Asia, Korea and Japan. Taiwanese manufacturers take the lead in TSV fabrication by CIS production for mobile-phones up to now.

1.2 Thermal Management Concepts and Limitations

1.2.1 Power Dissipation Characteristics of IC-Dies

Power delivery and heat removal to and from ICs became the limiting factors for a second time in history of electronic components [41]. It caused a paradigm change in technology and system architecture. Heat fluxes of of bipolar mainframe dies exceeded heat removal capabilities of cold plates in the late

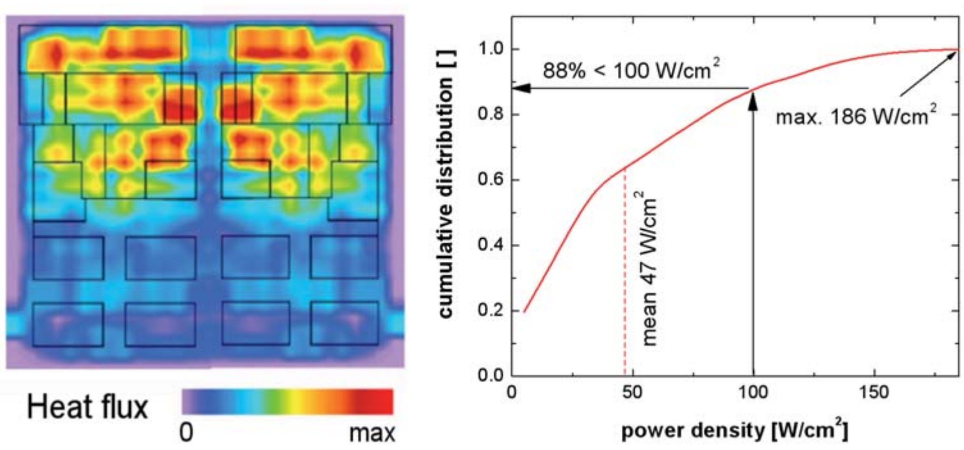

Figure 1.4: Measured power map of a duo-core processor at peak load using spatially-resolved infrared imaging [45] (left) (reprint with permission). cumulative distribution of the duo-core processor heat flux at peak load (right).

80ies. The industry was forced to change to energy efficient CMOS transistor technology and had to cope with a substantial cock-frequency drop. Fortunately, CMOS switches scale favorable and allowed ever increasing clock-frequencies at shrinking node dimensions in the following decade. The power dissipation at that times was caused primarily by the active power (P_{active}) of the transistors, which is a second-order function of the supply voltage (V_{dd}) and proportional to the gate capacitance (C) and clock frequency (f) [42]:

$$P_{active} \propto \frac{1}{2} C f V_{dd}^2. \tag{1.2}$$

The active power density turned out to be invariant down to 130 nm gate length even at increasing clock frequency. This, thanks to linear gate capacitance and supply voltage reduction with respect to the transistor physical dimension at constant-field scaling. Proportional V_{dd} to gate length scaling became impractical at smaller nodes due to parasitic effects. Passive power dissipation due to gate and subthreshold leakage became significant and account for nearly half the power dissipated in today's 45 nm devices [43, 44]. Due to the lack of alternative transistor technology the second paradigm change affected the microprocessor architecture. Clock frequency roadmaps to more than 10 GHz were abandoned and the multi-core era was entered. Total power consumption and clock-frequency are kept more or less constant. The microprocessor performance is now scaled by adding cores to the die. The next and third paradigm change will be a consequence of the multi-core architecture with it's increasing demand for cache size and bandwidth as discussed already in 1.1.1, leading to vertical integrated microprocessor - cache chip stacks in the near future.

The power map, displaying the spatial heat flux of a duo-core microprocessor was measured at peak load using spatially-resolved infrared imaging by HAMANN et al. [45] (Figure 1.4 left). High heat flux zones are caused by logic macros, compared to low power cache regions. The cumulative heat flux distribution is shown in Figure 1.4 (right). The peak heat flux of 186 $\frac{\mathrm{W}}{\mathrm{cm}^2}$ is close to four times the average heat flux (47 $\frac{\mathrm{W}}{\mathrm{cm}^2}$). It is defined as the total power dissipated divided by the chip area. In this example, only 12 % of the chip area dissipates more than 100 $\frac{\mathrm{W}}{\mathrm{cm}^2}$. This strong heat flux contrast is responsible for local hot-spots influencing signal delay and differential aging of transistors.

1.2.2 Thermal Response of IC-Dies

Temperature control of ICs is key to guarantee their reliability and to operate them at a defined freqeuency [41]. Charge carrier mobility and electrical conductivity of metal wires is reduced at elevated temperature. This alters the switching time of transistors and signal transmission delay. Furthermore, device efficiency is degraded by an exponential leakage current increase with temperature [46]. Hence, sub-ambient cooling utilizing refrigeration loops was proposed. The additional power dissipated by these compression loops with a coefficient-of-performance (COP) of about 3 to 5 exceeds the power sav-

Application	T_{jmax} [°C]
High-performance logic	95
Low-performance logic	125
Memory devices	125
Handheld devices	125
Automotive electronics	175

Table 1.2: Maximal junction temperature T_{jmax} for various applications according to ITRS.

ings from reduced junction temperature by far [47]. Therefore, the opposing trend to improve system efficiency by hot water cooling in conjunction with low temperature gradient microchannel cold plates can be observed. It allows to eliminate all refrigeration loops in the datacenter and is called free-cooling mode. The high quality output heat at a level of 65 °C can be sold to neighborhood-heating-networks where it's re-used for space heating purpose. It offsets the carbon emission and reduces total-cost-of-ownership of the datacenter [48]. Efficient computing becomes significant as world-wide power consumption of datacenters increases due to the consolidation trend of information and communication technology (ICT) [49].

Component life time (L) depends on aging effects, such as electro migration, diffusion, relaxation, delamination, and voiding. Most of these mechanisms depend exponentially on temperature (T) as described by the ARRHENIUS law:

$$L(T) = A\left(e^{\frac{E_a}{kT}} - 1\right) \tag{1.3}$$

with a system-specific constant A, the activation energy E_a and BOLTZMANN constant k [50]. The maximum junction temperature is then defined by the accepted mean-time-between-failure (MTBF) of the application. Typical industry standards defined by ITRS are listed in Table 1.2.

1.2.3 Established Heat Removal Concepts

Today's microprocessors are mounted on a wiring board using flip-chip technology. Individual solder balls with a spacing $\geq 150\,\mu\text{m}$ serve as electrical interconnects for signaling or power delivery purpose. This leaves the entire backside of the die for heat removal. A copper lid adhesively attached to the laminate spans the silicon die and acts as thermal conductive and mechanical protective interface from die to heat sink. The ubiquitous availability and compatibility of air with electronic components renders air cooling as the heat removal technique of choice. Air fins increase the total surface area and reduce the convective thermal resistance from solid to air by a factor of 100. A chip heat flux of up to 70 $\frac{\text{W}}{\text{cm}^2}$ can be dissipated. Higher heat fluxes can be removed with improved heat spreading in the heat sink base. Water- or methanol-filled heat pipes or vapor chambers are used accordingly. Fluid evaporates on top of the microprocessor die and condenses in the cold plate periphery with temperatures below the fluids boiling point. The fluid transport to the heat source is performed by capillary forces in wicking structures. A critical heat flux[6] of around 140 $\frac{\text{W}}{\text{cm}^2}$ constitutes the upper performance limit of such devices. Most efficient heat removal in terms of exergy destruction (low temperature gradient) and minimal pumping power is performed by microchannel cold plates utilizing water as a coolant. Heat flux levels up to 400 $\frac{\text{W}}{\text{cm}^2}$ could be demonstrated thanks to the high volumetric heat capacity of water which is 4183 $\frac{\text{kJ}}{\text{m}^3\,\text{K}}$, compared to air with 1.27 $\frac{\text{kJ}}{\text{m}^3\,\text{K}}$. The fluid flow in the microchannel cold plate was divided into six parallel flow sections populated with staggered fins. A cold plate thermal resistance of 7 $\frac{\text{K}\,\text{mm}^2}{\text{W}}$ was measured at a pressure of 0.35 bar and a volumetric flow rate of 1.7 $\frac{\text{l}}{\text{min}}$ [51]

Thermal interface materials account for efficient heat conduction between adjacent solid parts, such as the silicon die and the copper cold plate. Particle filled oils or adhesives at concentrations above the percolation threshold result in effective thermal conductivities of up to 6 $\frac{\text{W}}{\text{m}\,\text{K}}$. These materials form a

[6] All liquid is evaporated in the wicking structure at the critical heat flux causing a sudden drop in heat transfer. A thermal runaway of the system with catastrophic failure is the result.

compliant interface between CTE mismatched components, reducing thermo-mechanical stress during thermal cycling. Effective thermal resistances of 6 $\frac{\mathrm{K\,mm^2}}{\mathrm{W}}$ can be achieved [52].

1.2.4 Back-side Heat Removal Limits of Vertical Integrated Chip Stacks

Contemporary 2D thermal packaging technology will be used for 3D chip stacks initially, to mitigate development risk and to leverage on existing manufacturing infrastructure. Hence, the back-side heat removal potential needs to be evaluated (Figure 1.5 left). Heat fluxes of the active stratas are accumulating in the chip stack. Additional thermal barriers such as BEOL layers and bonding interfaces have to be considered. The arrangement, as well as the stacking sequence of high heat flux macros and dies for non-uniform in-plane and die-to-die power dissipation becomes an important design parameter [53, 41].

A first statement about scalability of back-side heat removal is possible with a compact thermal model assuming one-dimensional heat flux [54]. The chip stack is discretized into sub-components and is represented as equivalent resistor network with attached current sources (Figure 1.5 right). The linear equationsystem is solved using the GAUSS JORDAN elimination algorithm to derive individual junction temperatures T_{jn}.

The wiring layer thermal resistance R_b is calculated for a microprocessor and memory die considering BEOL dimensions for 45 nm node technology as reported by ITRS[7]. Effective-medium-theory is applied to define the effective thermal conductivity k_{eff_i} of individual wiring layers i. They are composed of copper k_{Cu} wires and vias with an area fill factor f and the inter-metal and inter-layer dielectrics k_{IMD} and k_{ILD}, respectively:

$$k_{eff} = f k_{Cu} + (1 - f) k_{ILD}. \tag{1.4}$$

The effective thermal resistance of individual BEOL layers is equal to the thickness t to effective thermal resistances ratio. Their sum

$$R_b = \sum_{i=1}^{n} \frac{t_i}{k_{eff_i}} \tag{1.5}$$

results in a thermal impedance of 1.43 $\frac{\mathrm{K\,mm^2}}{\mathrm{W}}$ and 3.50 $\frac{\mathrm{K\,mm^2}}{\mathrm{W}}$ for the memory and microprocessor wiring BEOL layers (Table 1.3).

A state-of-the-art thermal interface resistance of 6 $\frac{\mathrm{K\,mm^2}}{\mathrm{W}}$ and a cold-plate thermal resistance of 7 $\frac{\mathrm{K\,mm^2}}{\mathrm{W}}$ were used as reported in section 1.2.3. The thickness of the thinned dies was assumed to be 100 µm, compared to the 500 µm thick die interfacing the cold plate. Typical heat flux values (section 1.2.1) for memory stratas and microprocessor cache and logic (hot-spots) areas of 10, 60, and 240 $\frac{\mathrm{W}}{\mathrm{cm^2}}$ are assumed, respectively. The maximum acceptable junction temperature T_{jmax} of the microprocessor unit (MPU) is set to 80 °C at a fluid inlet temperature T_{fin} of 20 °C. The memory temperature limit is set to 95 °C. Critical temperatures for memory-logic and logic-logic stack configurations are calculated and depicted in Table 1.4 and visualized in Figure 1.6, respectively.

The temperature budget is not exceeded for a single MPU and multiple memory layers. MPU stacks of two dies can only be tolerated with misaligned high heat flux macros. In general, it is advantageous to place the high heat flux strata (MPU) close to the cold plate. Additionally, stacking of congruent high power dies should be prevented. It is questionable if these thermal design rules match with electrical needs. A large amount of TSVs on a regular grid have to be placed in the bottom dies to deliver the current to the top MPU[8]. This TSV grid causes a substantial silicon real-estate loss and constrains macro placement in the lower dies [55]. It demonstrates clearly - electrical and thermal design-rules are contradictory and trade-offs altering system performance are inevitable. Only marginal return-on-investments can be expected from 3D technology if thermal constraints are limiting the electrical design and therefore the device performance.

Hence, new, scalable heat-removal concepts have to be developed. In this context the proposal to include thermal vias on the die to improve heat conduction across large thermal barriers in the stack sounds like a drop in the ocean (besides adding additional routing congestion) [56].

[7] http://public.itrs.net

[8] Up to 3/4 of the electrical interconnects to a 2D MPU are today used to deliver power. This number will increase with further supply voltage reduction, unless on-chip step-down voltage conversion becomes feasible.

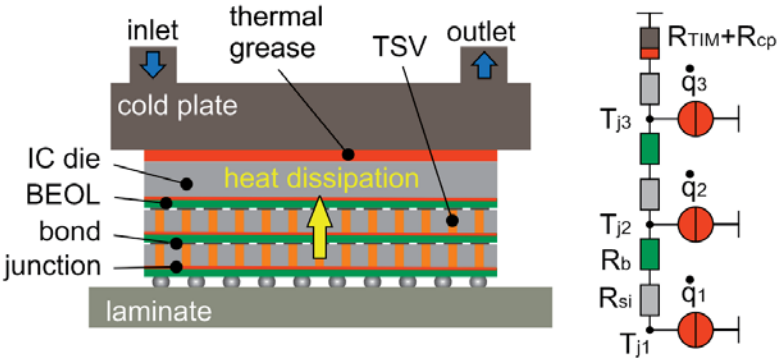

Figure 1.5: Back-side heat removal schematic illustrating chip stack with attached cold plate (left). Analogous resistor network representing a three-die stack with individual thermal resistances for the silicon substrate R_{Si}, the BEOL wiring layers R_b, the thermal interface R_{TIM}, and the cold plate R_{cp}. The heat flux is imposed by current sources $\dot{q_n}$ with resulting junction temperatures T_{jn} (right).

BEOL Layers	Microprocessor	Memory
Global / intermediate / metal 1 layers [count]	3x / 8x / 1x	1x / 2x / 1x
Total thickness [nm]	4180	2037
Effective thermal resistance [$\frac{\text{K mm}^2}{\text{W}}$]	3.499	1.434

Table 1.3: Number of wiring levels in the BEOL layers and its effective thermal resistance according to ITRS dimensions for a 45 nm technology node microprocessor and memory die.

MPU Budget:$\Delta T_{jmax-fin}$ =60 K	$\Delta T_{jmax-fin}$ [K]
11x memory / MPU logic / cold plate	59.0
MPU logic / 2x memory / cold plate	54.4
MPU logic / MPU cache / cold plate	60.8
2x MPU logic / cold plate	111.1

Table 1.4: Temperature gradient from fluid inlet to maximal junction temperature of the MPU ($\Delta T_{jmax-fin}$) for different chip stack configurations.

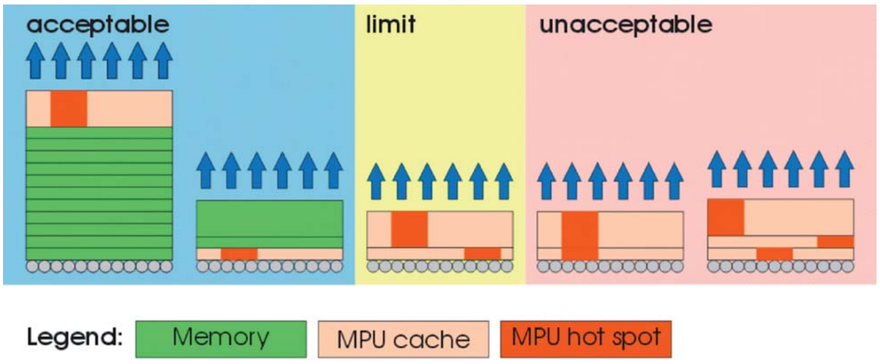

Figure 1.6: Chip stack configuration with acceptable, close to the limit and unacceptable junction temperatures for a 60 K thermal budget.

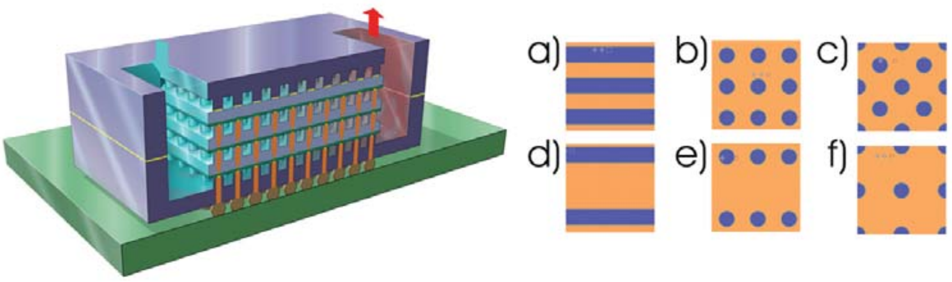

Figure 1.7: Artistic view of a interlayer cooled chip stack using forced convection (left). TSV-compatible heat transfer structures (blue: pin or fin), a) microchannel, b) pin-fin in-line, c) pin-fin staggered, d)-f) analogous structures, but half-populated (hp) instead of full-populated (fp) (right).

1.2.5 Interlayer Thermal Management

Volumetric heat removal concepts scalable with the number of stratas have to be investigated to keep 3D integration a successor of 2D MPUs with a enough room for improvement for several generations. This approach can potentially be performed by lateral heat removal between active dies relying on conduction or convection. Conductive layers between dies smooth out heat flux peaks and can dissipate the heat laterally out of the package. Layer thicknesses of more than 1 mm are needed to be thermally effective [57]. This limits the technology to peripherally-routed or low vertical interconnect density 3D packages and small die sizes (below 1 cm^2). Alternatively, thin form-factor vapor chambers may be implemented between stacked dies [58]. So far, the implementation of vertical-interconnects has not been demonstrated and the thermal performance was marginal. Both methods are not compatible with high interconnect densities and heat fluxes in large footprint chip size stacks. Only forced convection seams to be a feasible candidate for volumetric heat removal in high-performance vertically integrated chip stacks.

1.3 Convective Interlayer Heat Removal - a Scalable Concept

A coolant is forced by a pressure gradient through embedded fluid cavities between individual dies in the chip stack in case of convective interlayer cooling (Figure 1.7 left). The fluid cavities are defined by adjacent dies and solder ball arrays or fluid geometries etched into the backside of one of the silicon dies. The heat transfer structures need to be compatible with the area array arrangement of the electrical interconnects. Individual TSVs are integrated into micochannel walls or pins of pin-fin arrays (Figure 1.7 right). Dielectric coolants can be used without the need of electrical insulation, whereas hermetically sealing concepts are important to prevent hydrolysis or electrochemical corrosion if water is used.

Basic convective interlayer cooling was investigated by several researchers [59, 60, 61]. Takahashi demonstrated heat removal of 25 $\frac{\mathrm{W}}{\mathrm{cm}^2}$ for die stacks with peripheral interconnects and with water as coolant. Dies with a size of 10 mm formed a parallel-plate heat-transfer configuration at a spacing of 10 µm. Chen investigated dielectric coolants (FC-77) at a die spacing of 50 µm resulting in 50 $\frac{\mathrm{W}}{\mathrm{cm}^2}$ heat dissipation. Parallel-plate heat transfer geometries are not feasible for high-performance applications with the need for high density area array TSVs and the associated high heat-fluxes. Koo concluded with a critical heat flux of 140 $\frac{\mathrm{W}}{\mathrm{cm}^2}$ for a 2 cm^2 chip stack using two-phase heat transfer in microchannels and water as refrigerant. 300 µm high channels with a pitch of 800 µm were considered, resulting in a moderate TSV density. Integrated coolant delivery through the printed circuit board (PCB) was demonstrated by Dang et al. [62]. The fluid manifold consists of holes with a 50 µm diameter etched through the dies, resulting in a 2.5% silicon real-estate loss and a significant additional pressure drop. Buried microchannels were proposed. They were etched into the silicon die backside and where filled temporarily with a sacrificial spin-on polymer. Porous SiO$_2$ was deposited on top of the channels by