Pulsed Laser Annealing: A scalable and practical technology for monolithic 3D IC

Bipin Rajendran
Senior Member, IEEE
Dept. of Electrical Engg,
IIT Bombay, India
Email: bipin@ee.iitb.ac.in

Albert K. Henning
Senior Member, IEEE
MonolithIC 3D Inc, 3555
Woodford Dr, San Jose, CA
albertkhenning@yahoo.com

Brian Cronquist
Member, IEEE
MonolithIC 3D Inc, 3555
Woodford Dr, San Jose, CA
Email: briancronquist@att.net

Zvi Or-Bach
Member, IEEE
MonolithIC 3D Inc, 3555
Woodford Dr, San Jose, CA
zvi@monolithic3d.com

***Abstract*— Classical dimensional scaling faces challenges from growing on-chip interconnect time delays, and escalating lithography costs and layout limitations. In this paper, we present practical integration schemes for developing cost-efficient 3D ICs in a monolithic fashion, which employ fully depleted transistor channels and laser annealing to achieve sharper junction definition.**

Keywords—3D integration; SOI; laser annealing; simulation

I. INTRODUCTION

Classical dimensional scaling faces challenges from growing on-chip interconnect time delays and escalating lithography costs and layout limitations. Some have declared that the coming decade is the "3D Era" for both transistor shape and layer stacking. However, through silicon via (TSV) connected layer stacking is proving to be expensive. It is also relatively limited in vertical connection density (about $10^4/mm^2$), due to the large TSV size and keep-out-zone driven pitch [1]. In contrast, monolithic 3D provides 10,000 times higher vertical connection density (about $10^8/mm^2$) by means of sequential fabrication of thin (<100nm) layers [2]. Previous attempts at monolithic 3D integration have brought major compromises, whether to the overall device structure (by use of refractory metals) or the type of transistor used (for instance, TFT, RCAT [3]). These compromises are due to the thermal budget constraint (<400 °C) imposed by the existing underlying layer(s) on the processing of the layer above [4].

Two major semiconductor trends suggest how to develop a practical monolithic 3D IC process flow which uses conventional semiconductor processes and materials. The first major trend is the industry's move to fully depleted transistor channels (using bulk or SOI, FinFet or conventional MOSFET approaches), which requires much less silicon volume to fabricate the transistor channel [5]. The second major trend is the move to using shorter time-scale rapid thermal annealing, specifically the adoption of laser annealing to achieve a sharper junction definition [6]. We present a process flow leveraging an ion cut layer transfer process to obtain thin (<100nm) active layers of silicon. In order to build highly doped low leakage active regions in such thin layers, we show multiple options, including the use of pulsed excimer laser annealing (ELA) for dopant activation, and other important steps needed for formation of advanced transistors.

II. 3D-SYSTEM INTEGRATION CHALLENGES

A. Outline of major high temperature processes

One of the main challenges in monolithic integration of 3D ICs is the thermal budget: the processing time and temperature required for fabrication of good quality transistors in the upper layers could adversely affect the reliability of pre-existing circuitry in the lower layers. Table I shows the major high thermal budget process steps in a modern transistor fabrication process flow. We identify those process steps which are replaceable by low thermal budget based processes.

A portion of the steps in a semiconductor process flow, for example, a high speed CMOS FinFet or UTBB-FDSOI (Ultra Thin Body BOX – Fully Depleted Silicon On Insulator) transistor flow, may have processing temperatures which would damage the metallization (for example, copper or aluminum) and isolation layers (for example, low-k dielectrics, aerogels) in the substrate or layer below interconnect layers or regions. In addition, the processing temperature stress may change the device electrical characteristics and reliability of the devices (for example, transistors, capacitors, inductors, resistors) which reside in or on the substrate or layer below in the 3DIC stack.

Some of the higher temperature processing steps are described as follows: Si_3N_4 LPCVD depositions for STI (Shallow Trench Isolation) and transistor well formations are typically about 30 minutes at 700 °C, the STI liner oxidation is typically about 10 minutes at 800-1000 °C, and the STI TEOS/Flowable CVD densification is typically about 20 minutes at 1000 °C. Activation of the well implants is typically about 10 seconds at 1000 °C, formation of the dummy gate oxide is typically 10 minutes at 800 °C and the dummy gate amorphous silicon deposition is typically about 20 minutes at 600 °C. The raised source-drain/channel stressor SiGe and SiC/P selective epi is typically 10s of minutes at 600-700 °C and the silicide formation is typically minutes of RTA at 400-500 °C. Activation of the Source/Drain, halo, and Vt implants is typically a spike of seconds at 950 °C plus a flash or laser 3ms exposure at 1350 °C. The BIL (Bottom Interface Layer) oxidation and nitrogen stuffing is typically minutes at 825 °C

and the high-k material, such as HfO_2, post-ALD anneal is typically about 30 seconds at 700 °C. Overall, though, the highest temperature exposure is activation and damage annealing of the ion-implants.

TABLE I: MAJOR THERMAL PROCESS STEPS IN MODERN IC PROCESS

Step	Purpose	Thermal Budget	Replaceable?	With Laser?
Si_3N_4 LPCVD	STI and well	30 min – 700 °C	Yes	No
Liner Ox	STI	10 min – 800 °C	Yes	No
TEOS Densification	STI	20 min – 1000 °C	Yes	Yes
Implant Activation	Well	20 sec – 1000 °C	Yes	Yes
Dummy Ox	Gate	2 min – 800 °C	Yes	No
Dummy a-Si deposition	Gate	20 min – 600 °C	Yes	No
Selective epi deposition	SiGe and SiC S/D	20 min – 700 °C	Yes	No
Silicide formation	S/D contact	5 min – 400 °C	Yes	Yes
Implant Activation	S/D, halo, Vt	5 sec – 950 °C, LSA	No	Yes
BIL oxide+N	Gate	5 min – 825 °C	Yes	No
HfO_2 post ALD	Gate	30 sec – 700 °C	Yes	Yes

From the table above, it is clear that most critical high thermal budget processes required for a gate-last CMOSFET process could in principle be replaced or substituted by a short time scale process such as pulsed laser annealing.

B. Potential process flow

In order to fabricate a 3D Integrated Circuit in a monolithic fashion, conventional processes are used to fabricate the first layer of transistors on a silicon wafer. Additional shielding layers are deposited over this wafer and polished to create a smooth surface for bonding of active silicon through a layer-transfer process. The bonding surfaces are then subjected to clean and surface activation treatments (in O_2 plasma) to obtain strong, void free bonds [7]. The new c-Si layer from the donor wafer is separated using cleaving along an H^+ implant interface. This entire process is now becoming viable thanks to the fact that fully depleted transistors can be integrated atop these thin c-Si layers with relatively straight forward modifications of a gate-last CMOS process flow.

As listed in Table 1, most of the activation or anneal processes required for CMOS fabrication can be replaced with a pulsed laser annealing process. Further, low temperature alternate methods could be used in place of the traditional high thermal budget processes used for raised source drain formation and silicide contact formation.

An acceptor substrate/wafer may be conventionally processed including transistors, for example, MOSFETS, FD-MOSFETS, FinFets, and so on, as well as associated BEOL (Back End Of Line) interconnect copper wiring and inter-metal dielectrics. Thru layer via (TLV) metal interconnect strips or pads are also formed in the conventional BEOL processing. One or more (two are shown in Figure 1, below) shield/heat sink layers are formed with conventional BEOL processing, which may include copper or higher melting point materials such as tungsten. The top and bottom shield layer thickness shown in the example of Figure 3 are 250 nm of copper, and the inter-shield dielectric layers are 150 nm and 100 nm thick. The shield/heat sink layers have TLV path pads for pass thru of the vertical strata to strata signal connections, and alignment mark openings, and generally cover the majority of the die and scribe-lane area. The shield/heat sink layers may also have thermal paths/vias to the acceptor substrate constructed during the BEOL processing. The topmost shield layer is covered with a dielectric such as silicon oxide in preparation for bonding to the donor wafer.

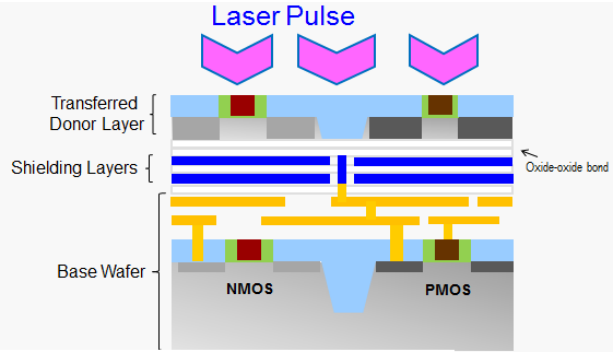

Figure 1: Schematic of an exemplary 3D-IC stack with shielding layers between stacked active layers, with laser annealing to activate the junctions in the upper active layers.

A monocrystalline donor wafer is prepared for layer transfer. One method is the well-known hydrogen ion-implant splitting [15, 16], or ion-cut, where a high dose of hydrogen ions creates an embrittled plane at a well-controlled and chosen depth of the donor wafer. The donor wafer is then flipped over and oxide-oxide bonded to the top of the acceptor substrate. Mechanical cleave, thermal cleave, or other layer transfer methods are utilized to remove the bulk of the donor wafer, thereby transferring a thin (10 to 100 nanometers) monocrystalline silicon layer on top of the acceptor substrate. The transferred layer may be protected by an oxide layer, shown as 5 nm thick in Figure 3 below. The bonding oxides are each 50 nm thick in the Figure 3 example.

The transferred donor layer may then be processed to form transistors with short pulsed laser exposures providing, for example, annealing of process damages and activation of dopants. It should be noted that the shield/heat sink layers are useful as Vss/Vdd planes or grids for the transistor layer above and may also serve as an emf shield between the transistor layers during chip operation.

III. SIMULATION OF LASER ANNEALING

A. Simulator design

We utilize a numerical simulator which includes temperature dependent material properties such as optical coefficients, thermal conductivity and specific heat of semiconductor materials to solve the light absorption and heat diffusion equations in a consistent manner to determine the absorption profile and thermal excursions for a representative 2D structure of the wafer stack. This wafer stack incorporates thermal via 'pipes' (whose effect is to increase the effective thermal conductivity of the annealed stack) and heat shielding schemes.

Structure to Model	Layer Name	Layer Number	Matl Number	Thickness	Units
	LTO1	1	1	0.1	um
	LTO2	2	1	0.1	um
	S/D2a	3	6	0.025	um
	S/D2b	4	6	0.025	um
	S/D2c	5	6	0.04	um
	S/D2d	6	6	0.01	um
	SiO2a	7	1	0.01	um
	SiO2b	8	1	0.19	um
	SiO2c	9	1	0.19	um
	SiO2d	10	1	0.01	um
	Gate1a	11	6	0.01	um
	Gate1b	12	6	0.19	um
	Gate1c	13	6	0.1	um
	Gate1d	14	6	0.1	um
	Gate1e	15	6	0.1	um
	Gate Oxide 1a	16	1	0.025	um
	Gate Oxide 1b	17	1	0.025	um
	Substrate-a	18	2	0.025	um
	Substrate-b	19	2	0.05	um
	Substrate-c	20	2	0.125	um
	Substrate-d	21	2	0.8	um
	Substrate-e	22	2	9	um

Figure 2: 1-D model of the exemplary 3D IC stack used for simulation benchmarking. In this stack, green is SiO_2, black is amorphous Si (a-Si), and gray is single-crystal Si (c-Si). Material properties, including temperature dependences, are found in Refs. [9-11].

For this work, a quasi-lumped element approach is taken to model the time-dependent evolution of temperature in relevant 3D IC layers. The model is 'quasi', insofar as the time steps are taken to be quite small relative to the thermal diffusion time in the various material layers. However, the layer thicknesses are allowed to be somewhat thick relative to the thermal diffusion length. To overcome the possibility of unrealistic transient results, layers which are thicker than the thermal diffusion length are broken up into sub-layers. Said another way, the flow of heat in any particular sub-layer is taken to be in steady-state at each simulation time-step, so that the slope of thermal flux vs. distance throughout each sub-layer is a constant. But, because layers are broken into sub-layers, the overall flux through the entire layer need not be a constant of space at any particular instant in time.

The lumped-element model is implemented using Microsoft Excel, and the Visual Basic programming language which underpins Excel. The generic set of 1D model equations is as follows:

$$dT_1 = \frac{dt}{m_1 C_1}\left\{\frac{T_{TOP}-T_1}{R_{TOP\_1}} - \frac{T_1-T_2}{R_{1\_2}}\right\}$$

...

$$dT_i = \frac{dt}{m_i C_i}\left\{\frac{T_{i-1}-T_i}{R_{i-1\_i}} - \frac{T_i-T_{i+1}}{R_{i\_i+1}}\right\} \qquad (1)$$

...

$$dT_n = \frac{dt}{m_n C_n}\left\{\frac{T_{n-1}-T_n}{R_{n-1\_n}} - \frac{T_n-T_{BOT}}{R_{n\_BOT}}\right\}$$

Boundary conditions are set at the TOP and BOT locations in the 1D stack. The state variables are the sub-layer temperatures T_i. m_i, C_i and R_i are the mass, heat capacity and thermal resistance, respectively, for each sub-layer. dt is made small enough so that convergence and stability are obtained. Because material parameters for density, specific heat, and thermal conductivity are inherent in m_i, C_i and R_i, and because these material parameters are temperature-dependent, the system of equations is solved without resort to matrix inversion.

Initial conditions for the state variables are set at time t=0. The system solution then proceeds forward in time, until either the end-point of the simulation is reached, or steady-state throughout the system is achieved.

To benchmark the model, we turn to the earlier work found in Figures 4.1 and 4.4 of Ref. [8]. Figure 2 shows the sub-layer structure, cast in the model of Equations (1).

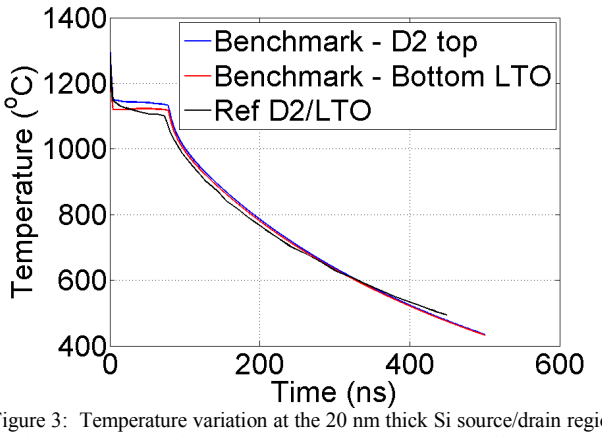

Figure 3: Temperature variation at the 20 nm thick Si source/drain region in the upper active layer during laser annealing. Excellent match is obtained between our quasi-lumped approach compared with the finite element solver used in the reference [8].

Because we are interested in the use of pulsed laser annealing to drive silicon layers beyond the melting point,

following the suggestion of [13] an accommodation for the heat of fusion was made, in the temperature-dependent equation for the specific heat of silicon. In this instance, in MKS units the specific heat becomes:

$$C_p(T) = \left(\frac{1000}{28.085}\right)\left[23.698 + 3.305\cdot 10^{-3}T - 4.354\cdot 10^{-5}T^{-2}\right]$$

$$+\left(\frac{1.788\cdot 10^{6}}{\delta T\sqrt{2\pi}}\exp\left[-\frac{(T-T_m)^2}{2\cdot \delta T^2}\right]\right) \qquad (2)$$

The exponential term accommodates the phase transition. The width of the transition in temperature is modeled by the factor δT. If phase change in the silicon is important, then an indirect constraint is that the time steps must be small enough, such that the change in temperature dT for a silicon sub-layer must be a fraction of δT through the phase transition.

Figure 3 compares the output from the quasi-lumped-element model, with the full partial differential equation-based results for the structure in Figure 1 given in [8].

Given the success of the benchmark, we proceeded to explore the viability of the use of copper shield layers in monolithic 3D integrated circuit processing and construction. The purpose of these layers is to ensure the underlying circuits, already made, do not suffer inordinately high temperature excursions, even when the newly-transferred surface silicon layer is driven beyond the solid-liquid phase transition. In particular, we sought to ensure the underlying structures would not experience a thermal excursion, at any time during the thermal processing, of an absolute temperature greater than 200 °C.

Structure to Model	Layer Name	Layer Number	Matl Number	Thickness	Units
	Protect	1	1	0.005	um
	Transferred Layer	2	2	0.02	um
	Donor Bond Ox	3	1	0.05	um
	Acceptor Bond Ox	4	1	0.05	um
	Top Shield	5	3	0.25	um
	Inter-Shield	6	1	0.15	um
	Bottom Shield	7	3	0.25	um
	Inter-Shield	8	1	0.1	um
	IC8	9	3	0.1	um
	Ins8	10	4	0.09	um
	IC7	11	3	0.1	um
	Ins7	12	4	0.09	um
	IC6	13	3	0.1	um
	Ins6	14	4	0.09	um
	IC5	15	3	0.1	um
	Ins5	16	4	0.09	um
	IC4	17	3	0.1	um
	Ins4	18	4	0.09	um
	IC3	19	3	0.1	um
	Ins3	20	4	0.09	um
	IC2	21	3	0.1	um
	Ins2	22	4	0.09	um
	IC1	23	3	0.1	um
	Ins1	24	4	0.09	um
	IC0	25	3	0.05	um
	Ins0	26	4	0.05	um
	Substrate Si	27	2	775	um

Figure 4: 1-D model of the exemplary 3D IC stack comprising of 2 shield layers separating 8 level metallization of lower level devices in a wafer and the transferred layer that is being subjected to laser annealing.

The 1D structure was set up according to Figure 4. Eight layers of metallization are assumed. Two copper shield layers are employed, with thicknesses as shown (derived from ITRS 20 nm 2013 MPU table). Insulator layers in yellow are taken to be SiCO. The system is modeled where time t=0 is immediately after a pulsed laser annealing step. The laser energy is predominantly, but not completely, deposited in the surface silicon layer. The assumption is that the surface silicon layer reaches a temperature of 1440 °C; to reach this temperature, a 20 nm c-Si layer requires (in the model) 14.68 mJ/cm^2 of laser power. Absorption coefficients for the copper layers were found in Ref. [14] from the Filmetrics web site. For a wavelength of 308 nm, the Cu absorption coefficient is 0.12 nm^{-1}. As a consequence, the top Cu shield starts the simulation at 123 °C, while the bottom Cu shield starts the simulation at 31 °C.

Figure 5a shows the results of the simulation, detailing only the surface (transferred) Si layer, the top Cu shield, and the bottom Cu shield. Figure 5b shows a closer inspection of the transient response. The top shield's temperature does not exceed 185 °C, while the bottom shield does not exceed 85 °C. The surface Si layer solidifies in just under 2 ns.

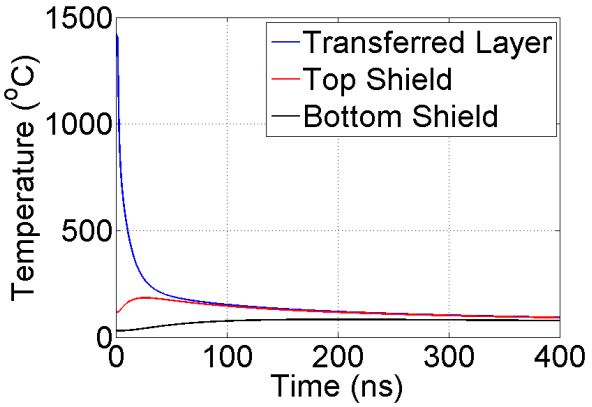

Figure 5a: Temperature excursion calculated using the electro-thermal simulation framework for the 1-D model of the exemplary 3D IC stack shows that the transferred layer can be heated above 1200 °C.

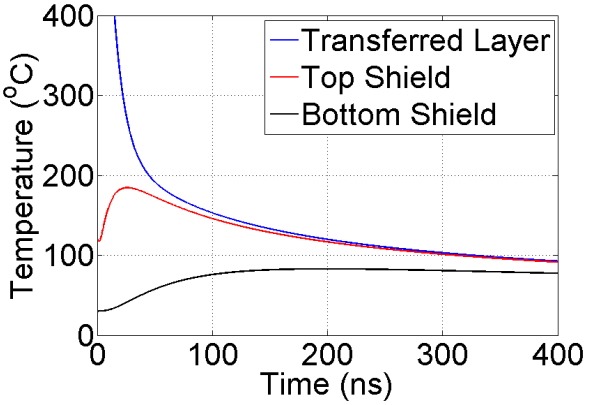

Figure 5b: Temperature excursion calculated using the electro-thermal simulation framework for the 1-D model of the exemplary 3D IC stack shows that the maximum temperature attained by the shield layer is limited to below 200 °C.

Conceptually, this result may initially appear to be counter-intuitive: the top Si is driven by the pulsed laser anneal into the melt regime, and there is no laterally thermal conductivity assumed for the Cu shield layers, to remove heat and protect the underlying IC structures already built. However, two factors emerge which are critical to understanding the success of this structure design for pulsed laser annealing processes. First, the transferred Si layer is very thin, 20 nm. As a consequence, the amount of energy which much be dissipated is actually quite low, even though the surface Si layer begins in the liquid state. Second, the heat capacity of the Cu shields is more than sufficient to absorb the energy dissipated by the surface Si layer as it cools; no additional thermal shunting laterally, using thermal conductivity as opposed to heat capacity in the Cu shields, is necessary.

IV. CONCLUSIONS

Using numerical simulations based on a lumped element approach (which has been validated against accurate FEM calculations), we show that as the thickness of the transistor silicon that needs to be annealed or activated approaches 20 nm, pulsed laser annealing -- for example, with pulse widths in the sub-100ns time scales and wavelengths in the 200-350 nm range -- can momentarily heat the top layer to > 1400 ^{0}C while the underlying device layers are kept below 150 °C. Thin shielding layers are used to achieve the desired thermal isolation between the bonded active layer at the top and the interconnect layers at the lower level. Pattern density effects associated with non-uniform absorption of laser light at these advanced technology nodes can be mitigated by using absorption layers or by employing a close-to-Brewster angle laser incidence. These results suggest a wide range of known and available rapid annealing techniques and equipment choices for enabling monolithic 3D integration are now practical.

REFERENCES

[1] J. H. Lau, "TSV manufacturing yield and hidden costs for 3D IC integration," Proceedings, 60th Electronic Components and Technology Conference (ECTC),.pp. 1031-1042 (2010).

[2] A. W. Topol, *et al*., "Three-dimensional integrated circuits," IBM Journal of Research and Development , vol. 50, no. 4.5, pp.491-506, (2006).

[3] C. S. Deepak, *et al*., "Monolithic 3D-ICs with Single crystal Silicon Layers", IEEE 3DIC Conference, 2012.

[4] B. Rajendran, *et al*., "Low Thermal Budget Processing for Sequential 3-D IC Fabrication," IEEE Trans. Elec. Dev.,.vol. 54, no. 4, pp.707-714 (2007).

[5] P. Flatresse, *et al*., "Ultra-wide body-bias range LDPC decoder in 28nm UTBB FDSOI technology," Digest of Technical Papers, IEEE Intl. Solid-State Circuits Conference, pp. 424-425 (2013).

[6] E. M. Bazizi, *et al*., "Analysis of USJ Formation with Combined RTA/Laser Annealing Conditions for 28nm High-K/Metal Gate CMOS technology Using Advanced TCAD for Process and Device Simulation," Silicon-Germanium Technology and Device Meeting (ISTDM), pp.1-2 (2012).

[7] X. X. Zhang and J.-P. Raskin, "Low-Temperature Wafer Bonding: A Study of Void Formation and Influence on Bonding Strength," IEEE Journal of Microelectromechanical Systems, vol. 14, no. 2, (2005).

[8] B. Rajendran, "Low thermal budget processing for sequential three dimensional integrated circuit fabrication." Ph.D. dissertation, Stanford University (August 2006).

[9] H. F. Wolf, "Silicon semiconductor data." Permagon Press, Oxford (1969).

[10] C. J. Glassbrenner and G. A. Slack, "Thermal conductivity of silicon and germanium from 3 K to the melting point," Phys. Rev. 134, pp. A1058-A1069 (1964).

[11] V. M. Glazov and A. S. Pashinkin, "The thermophysical properties (heat capacity and thermal expansion) of single-crystal silicon," High Temperature 39, pp. 413-419 (2001).

[12] H. Wada and T. Kamijoh, "Thermal conductivity of amorphous silicon," Jap. J. Appl. Phys. 35, pp. L648-L650 (1996).

[13] W. Ogoh and D. Groulx. "Stefan's Problem: validation of a one-dimensional solid-liquid phase change heat transfer process." In Proceedings, COMSOL Conference (2010).

[14] H.-J. Hagemann, W. Gudat, and C. Kunz, "Optical constants from the far infrared to the x-ray region: Mg, Al, Cu, Ag, Au, Bi, C, and A12O3," JOSA 65, pp. 742-744 (1975)

[15] M. Bruel, "Process for the Production of Thin Semiconductor Films", U.S. Patent No. 5,374,564, Dec. 20, 1994.

[16] M. Sadaka, et al., Building Blocks for Wafer-Level 3D Integration, Solid State Technology, August 18, 2010