Supplementary Information

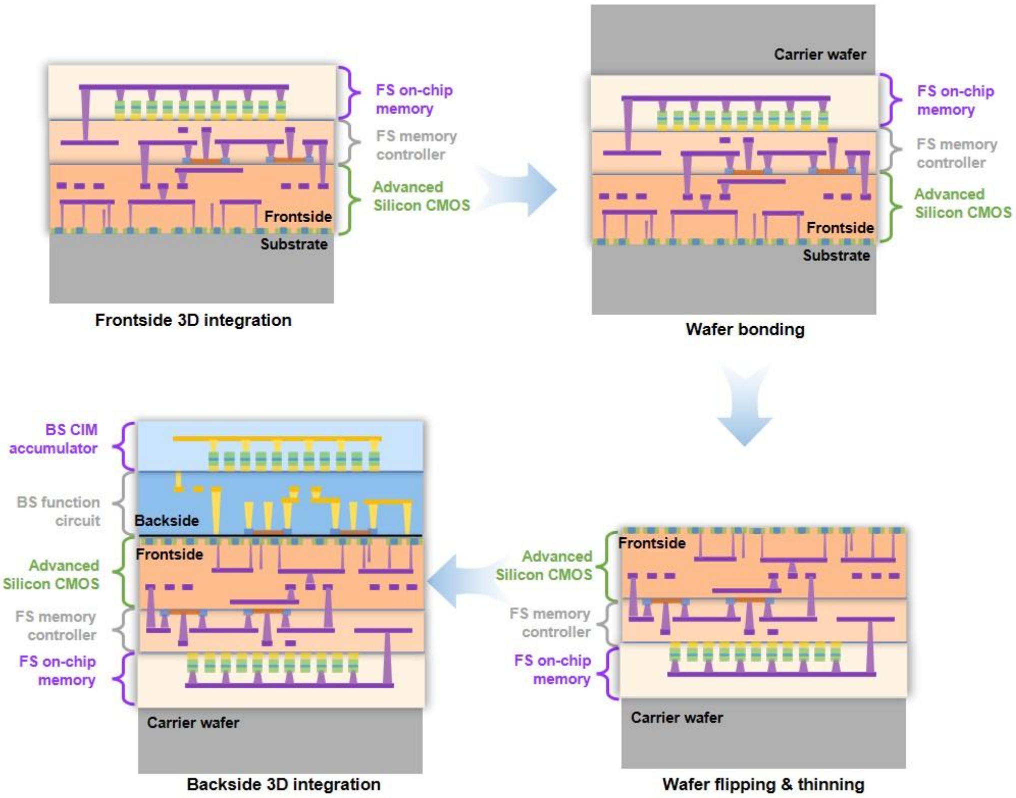
Unifying Heterogeneous and Monolithic Integration via Dual-sided 3D Technology

Yanbang Chu[1], Peiyan Hong[2], Jianxiang Jin[1], Jingwei Sun[1], Gaoqi Yang[1], Rui Guo[1], Cong Li[1], Zhentao Xiao[1], Ze Liu[1], Chenhao Xue[1], Yandong Ge[1], Yu Li[3], Xiuzhen Li[4], Chenyang Cui[4], Jundong Zhu[4], Liyan Dai[4], Lining Zhang[3], Fengwen Mu[5], Jin Kang[1], Weihai Bu[1], Zongwei Wang[1], Xuefei Li[2], Yimao Cai[1], Xin Zhang[1], Guangyu Zhang[4], Yanqing Wu[1], Guangyu Sun[1], Ming Li[1], Runsheng Wang[1,*], Heng Wu[1,*], Ru Huang[1]

[1]School of Integrated Circuits, Peking University, Beijing, China
[2]Wuhan National High Magnetic Field Center and School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan 430074, China
[3]School of ECE, Peking University, Shenzhen, China
[4]Beijing National Laboratory for Condensed Matter Physics and Institute of Physics, Chinese Academy of Sciences, Beijing, China
[5]iSABers Group Co., Ltd. Tianjin, China
[*] wrs@pku.edu.cn, hengwu@pku.edu.cn

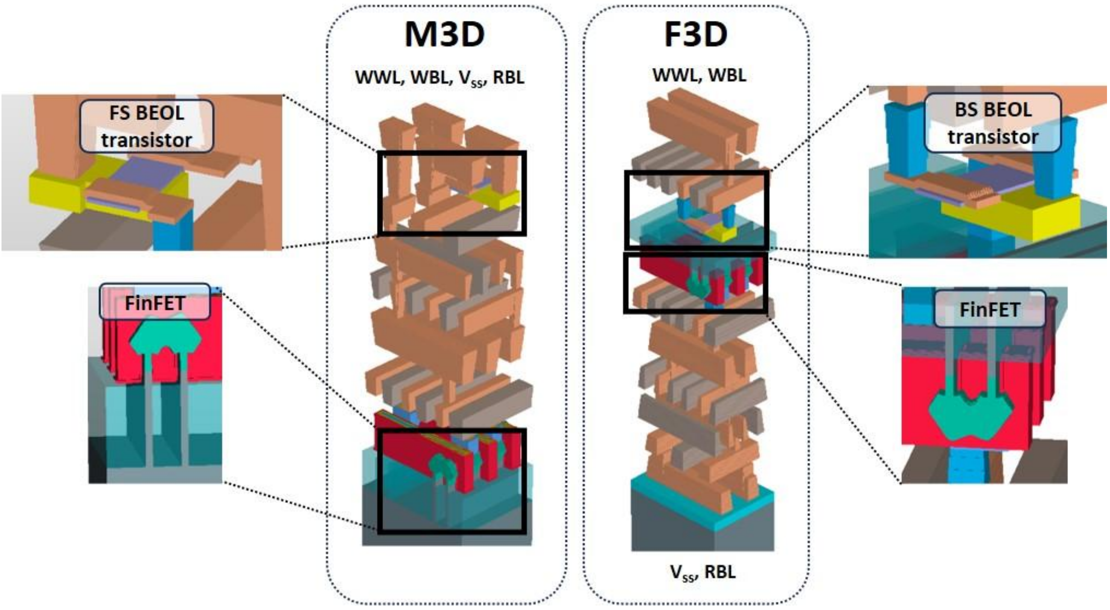
Table of Contents

Supplementary Note 1. Basic process flow of F3D scheme

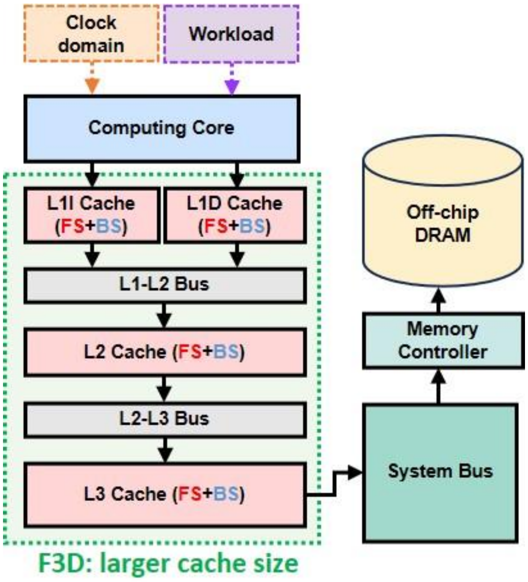


Supplementary Figure 1. The process flow of F3D, enabled by wafer bonding and flipping. After the frontside function layers integration, a carrier wafer is bonded onto the upmost surface of wafer, then the wafer is flipped and the substrate is thinned by grinding and CMP to the very bottom of the frontside transistor. At last, the connectivity between frontside and backside can be fabricated and more function layers can be integrated on the backside of wafer.

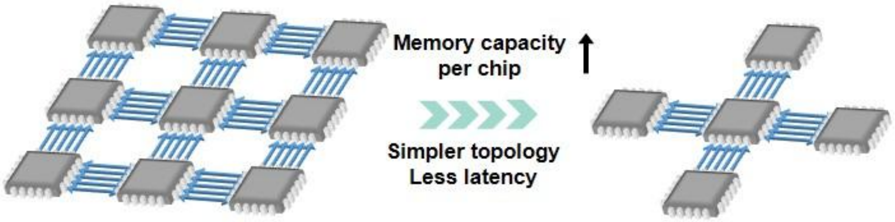
Supplementary Note 2. Evaluation of F3D at circuit level and system level



Supplementary Figure 2. The F3D benefits in the circuit level by 3D structure demonstration. Taking 2T0C DRAM as an example, F3D features shorter connection distance between BEOL transistor and FinFET compared to M3D. Moreover, due to the back-to-back stacking nature, the circuit pins (WWL, WBL, V_{ss}, RBL as indicated in the middle figure) in F3D can be placed on the both sides of wafer, resulting in much more relaxed routing congestion (the dark orange metal wires indicate the routing wires used for signal routing, the gray metal wires indicate dummy metal wires). The relaxed routing congestion will optimize the parasitic and improve the performance of circuit, which is discussed in the main text.

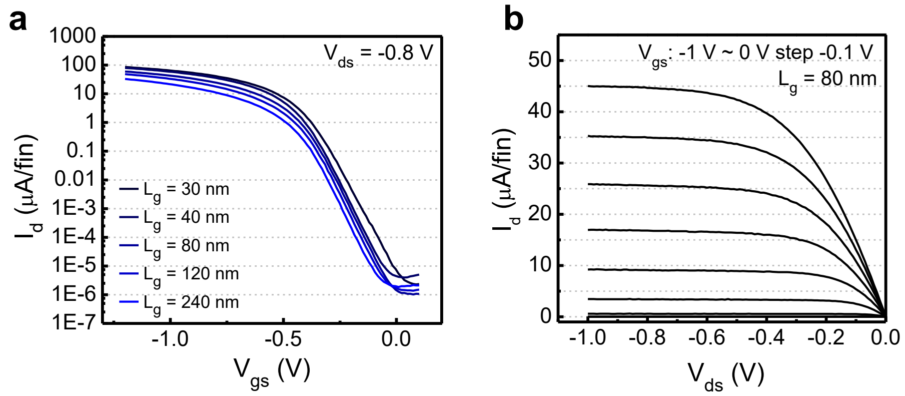


Supplementary Figure 3. Architecture of the CPU system used for performance simulation. The computing core deals with the input workload, and the cache system stores the temporary data or instructions with high required rates. The F3D can provide more cache memory size with less latency, as indicated by the green dash line box. This will bring gains in workload host time and cache miss rates, as quantified in Fig. 1d. Note that, the latency benefits are not considered in the simulation.

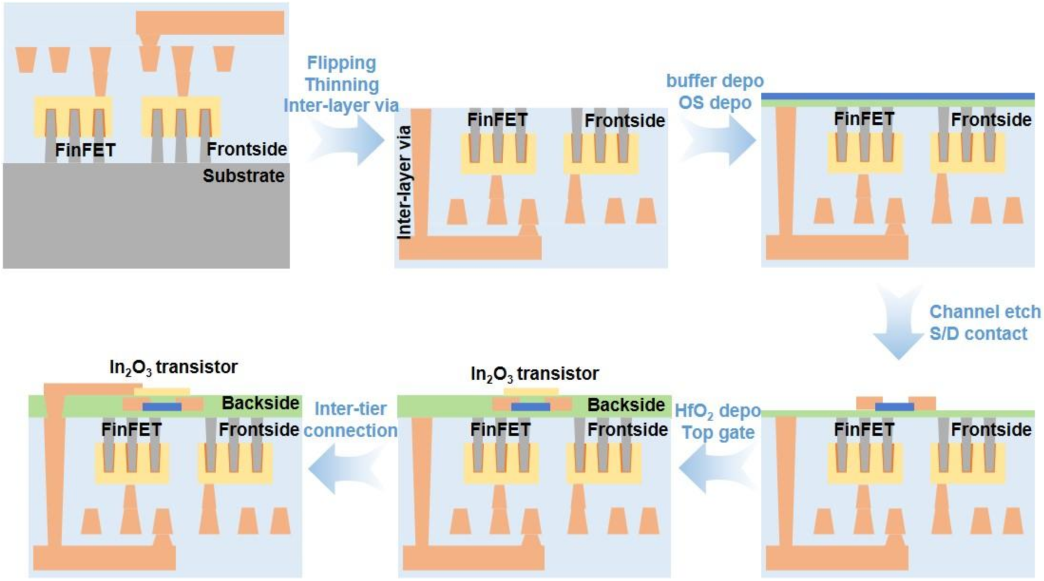


Supplementary Figure 4. The benefits in AI module deployment in F3D. By utilizing the wafer backside for on-chip main memory placement, larger memory capacity and less chips are needed for the parameter storage. Therefore, simpler interconnect network topology, smaller data transfer latency and power consumption can be obtained.

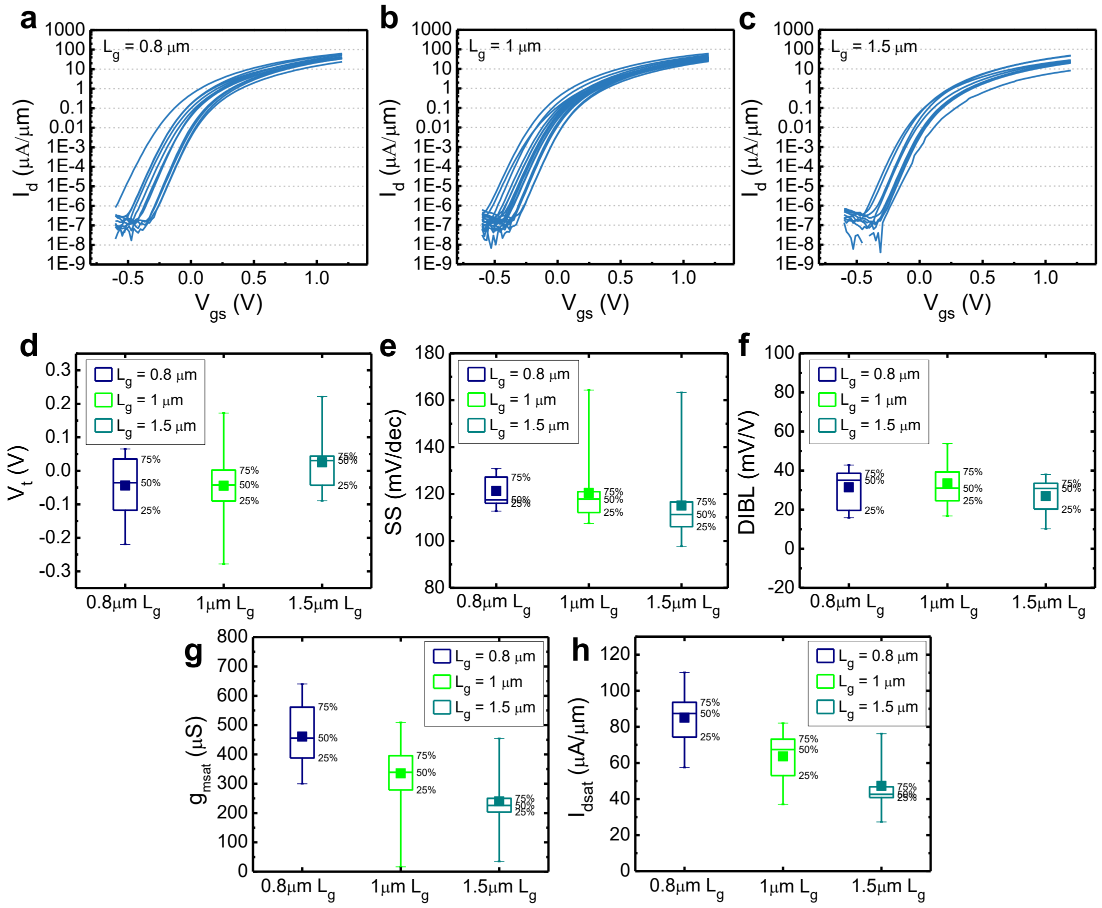
Supplementary Note 3. Characteristics of frontside FinFET and backside OS transistor



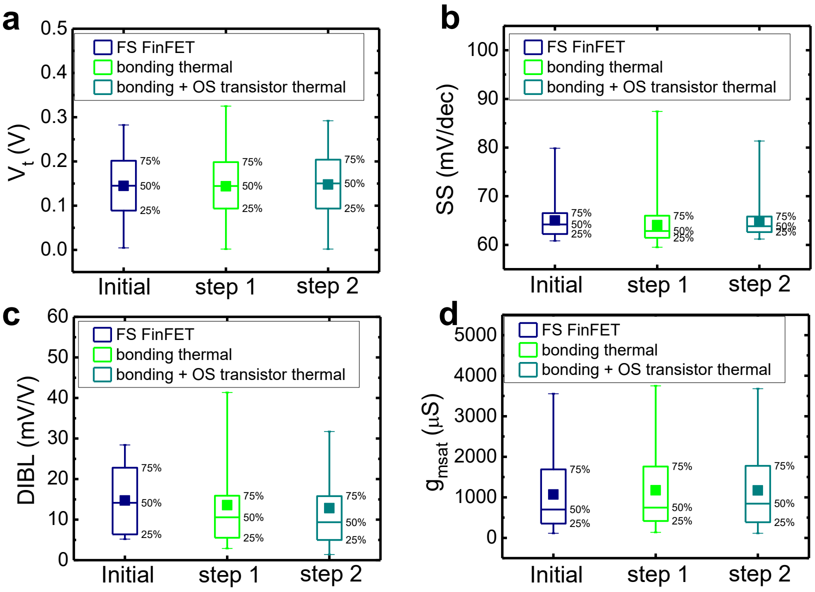
Supplementary Figure 5. The electrical characteristics of standard FinFET in this paper. (a) Transfer curves with varied gate length (L_g). The drain current is normalized by the fin number. **(b)** Output curves of FinFET with L_g = 80 nm.



Supplementary Figure 6. The process flow of BS OS transistor integration with FS FinFET. After the wafer flipping and substrate stripping, the inter-layer via is first formed. Then, the buffer layer and In_2O_3 channel layer are deposited, followed by the OS transistor formation. At last, the frontside to backside connection is formed through the HfO_2 layer.

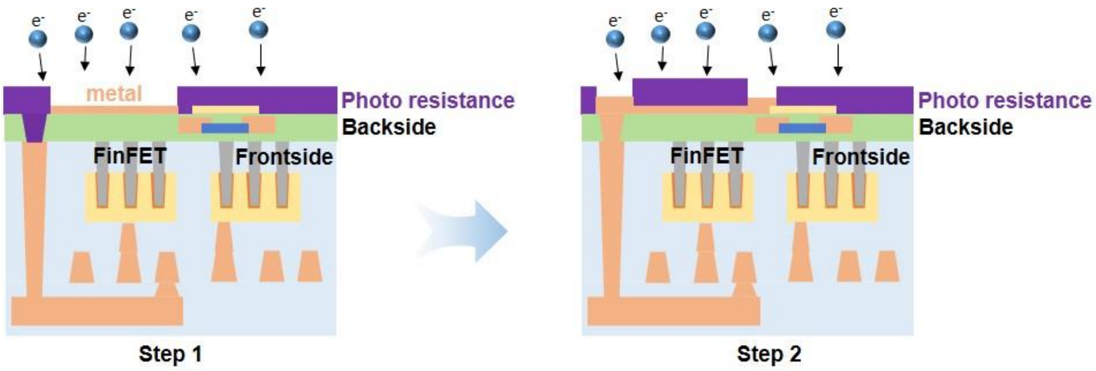


Supplementary Figure 7. The electric characteristic of backside OS transistors. (a-c) Transfer curves of OS transistor with gate length L_g = 0.8 μm (a), 1 μm (b), 1.5 μm (c). **(d-h)** Box plots of key transistor performance metrics with varied L_g for Vt, SS, DIBL, g_{msat} and I_{dsat}. With scaled L_g, OS transistors hold stronger drivability with the cost of more negative Vt and degraded SS.



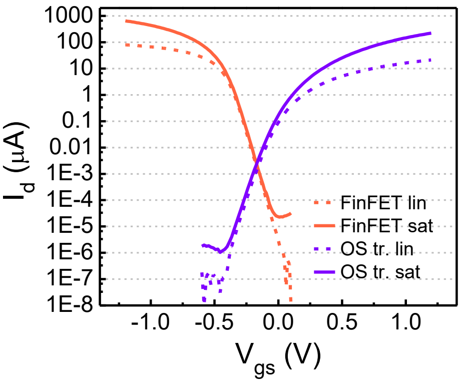
Supplementary Figure 8. The impacts on FinFET of F3D process thermal budget. The key thermal budget study is separated into two stages: the wafer bonding thermals (indicated as step 1) and OS transistor formation thermals (indicated as step 2). It shows negligible influence on the FinFET in the aspects of short channel effect and drivability, validating the F3D scheme's compatibility with advanced silicon transistor.

Supplementary Note 4. Heterogeneous CMOS inter-layer connect methodology

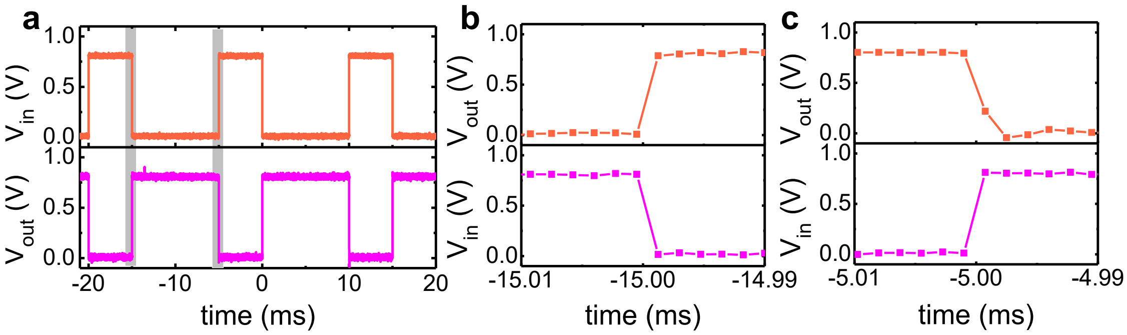


Supplementary Figure 9. The jumper wire method to protect the frontside FinFET during backside processes. For step 1, the interconnect metal is placed close to FinFETs' S/D/G terminals but not connected to avoid the antenna effect. At step 2, the jumper wire is formed to finalize the interconnect with small exposed area. Thanks to this, robust interconnects between the frontside and backside are formed with minimized antenna effect.

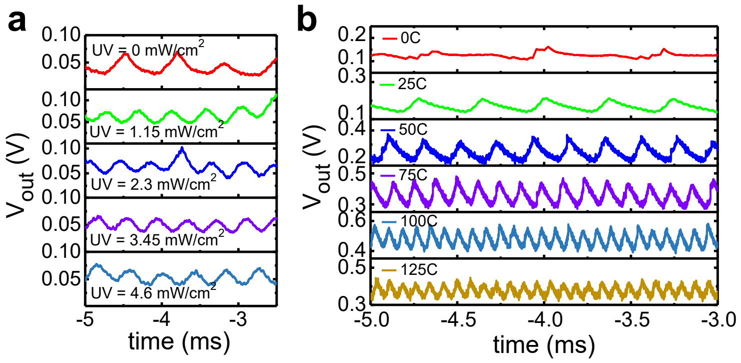
Supplementary Note 5. Heterogeneous CMOS circuit characteristics



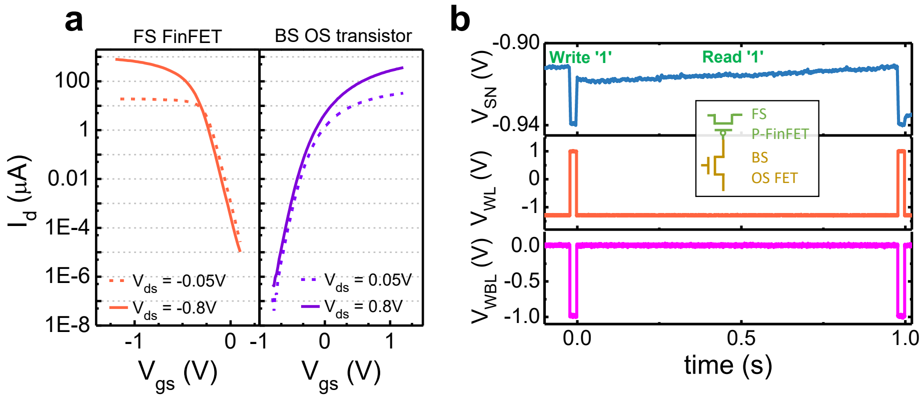
Supplementary Figure 10. The transfer curves of frontside FinFET and backside OS transistor (OS tr.) for heterogeneous CMOS. The relative negative threshold voltage of OS transistor (~ -0.25 V) compensates the slightly lower drivability of OS transistor, making the heterogeneous inverter's switching threshold voltage around $V_{dd}/2$. The "lin" and "sat" indicate linear regime (V_{ds} = 0.05 V) and saturation regime (V_{ds} = 0.8 V), respectively.



Supplementary Figure 11. The dynamic behavior of heterogeneous inverters based on F3D. (a) The waveforms of a heterogeneous inverter in response to an input square waveform. Stable output switching can be achieved. The gray boxes indicate the zoom-in range for (b-c). **(b)** The zoom-in waveform of V_{out} fall transition from (a). **(c)** The zoom-in waveform of V_{out} rise transition from (a). Balanced rise and fall transition time can be seen. The relatively large transition time is resulted from the large load of oscilloscope for measurement.

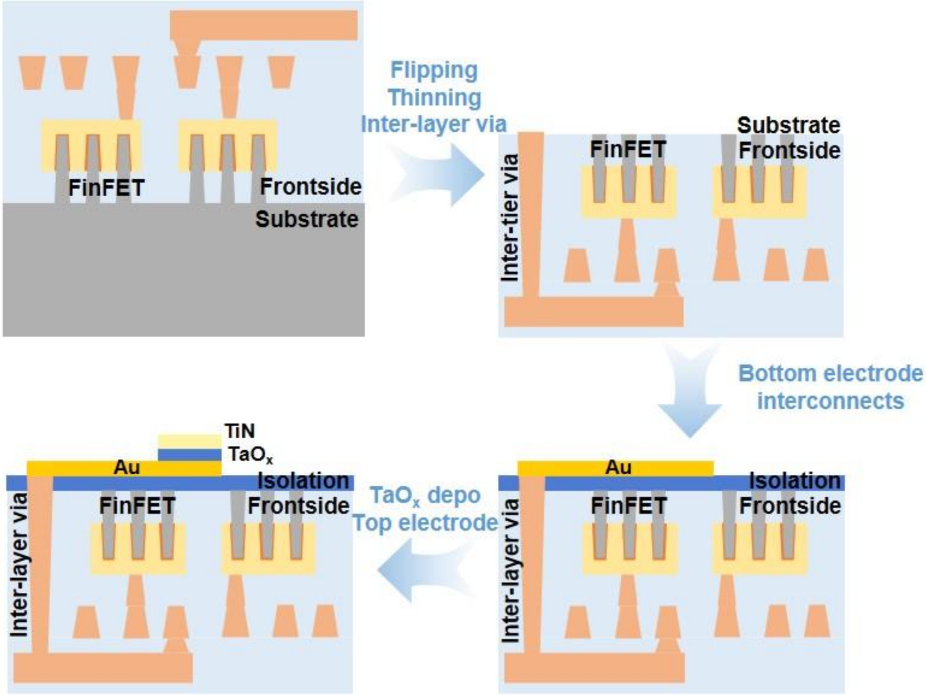


Supplementary Figure 12. The frequency response of the heterogeneous RO for different environmental parameters. (a) Waveforms of RO under UV illumination with various power. **(b)** Waveforms of RO at various working temperatures.

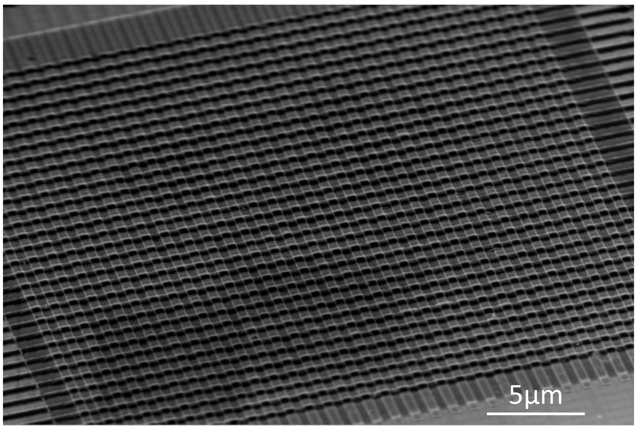


Supplementary Figure 13. The bipolar 2T0C DRAM demonstration in the F3D form. (a) Transfer curves of PMOS FinFET as the read transistor (left) and BS OS NMOS transistor as the access transistor (right). **(b)** The function waveform of bipolar 2T0C DRAM, indicating longer retention time than unipolar 2T0C DRAM.

Supplementary Note 6. Backside RRAM demonstration

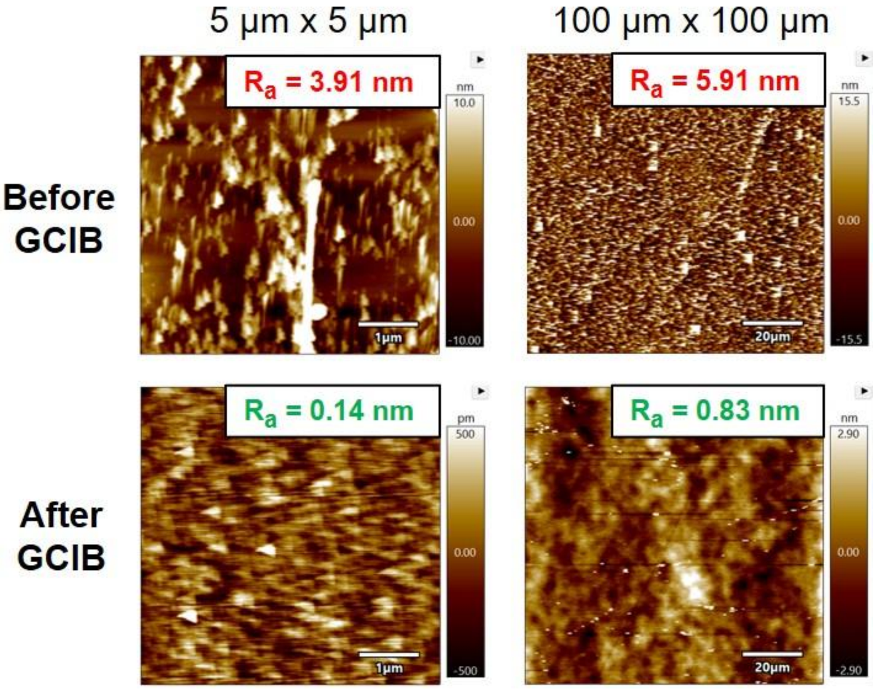


Supplementary Figure 14. The process flow of BS RRAM integration. After wafer flipping and substrate thinning, the inter-layer via is formed. Then the gold bottom electrode (connected to the inter-layer via), TaO_x dielectric layer and TiN top electrode are formed in sequence to fabricate the BS RRAM.

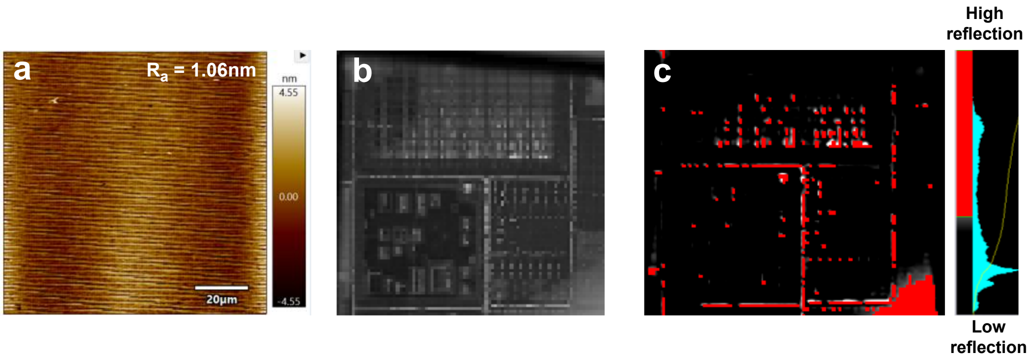


Supplementary Figure 15. The SEM image of RRAM array formed at the backside of wafer. The array size is 32×32 and the RRAM bitcell size is 300 nm × 400 nm.

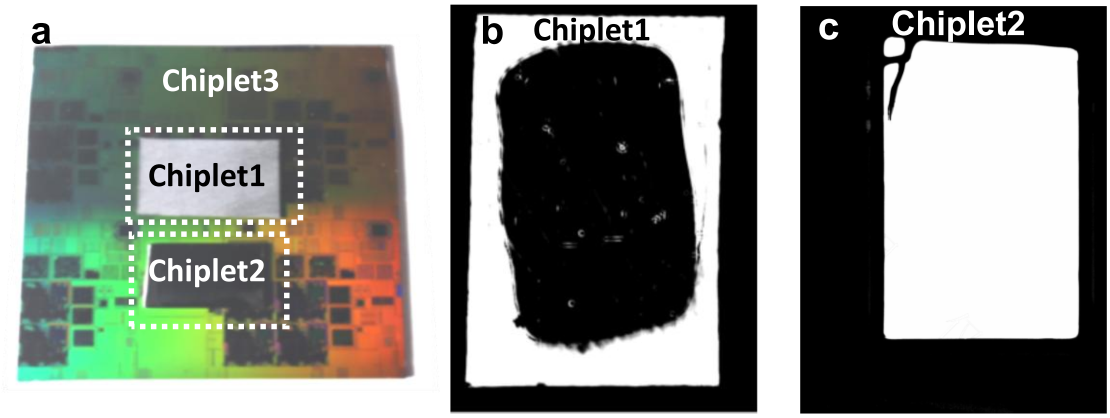
Supplementary Note 7. Dual-sided heterogeneous bonding technology development



Supplementary Figure 16. The bonding surface flattening by gas cluster ion beam (GCIB) etch. The top two figures show the AFM images before GCIB flattening. The bottom two figures show the AFM images of the same sample after the GCIB treatment. Clear roughness reduction of over one order of magnitude can be seen.



Supplementary Figure 17. The bonding quality assessment of inter-chip bonding. (a) The optimized surface roughness of < 1.1 nm before bonding. **(b-c)** The scanning acoustic microscope image of the chip after bonding. The circuit embedded in the silicon can be clearly seen, meaning the great bonding quality without bubbles. Black region in (c) indicates the area with low acoustic reflection (or high bonding quality).



Supplementary Figure 18. The bonding quality of the dual-sided chiplet bonding sample. (a) The optical image of the sample. The white dash boxes mark the region imaged by SAM in (b-c). **(b)** The SAM image of bonded chiplet1. **(c)** The SAM image of bonded chiplet2.