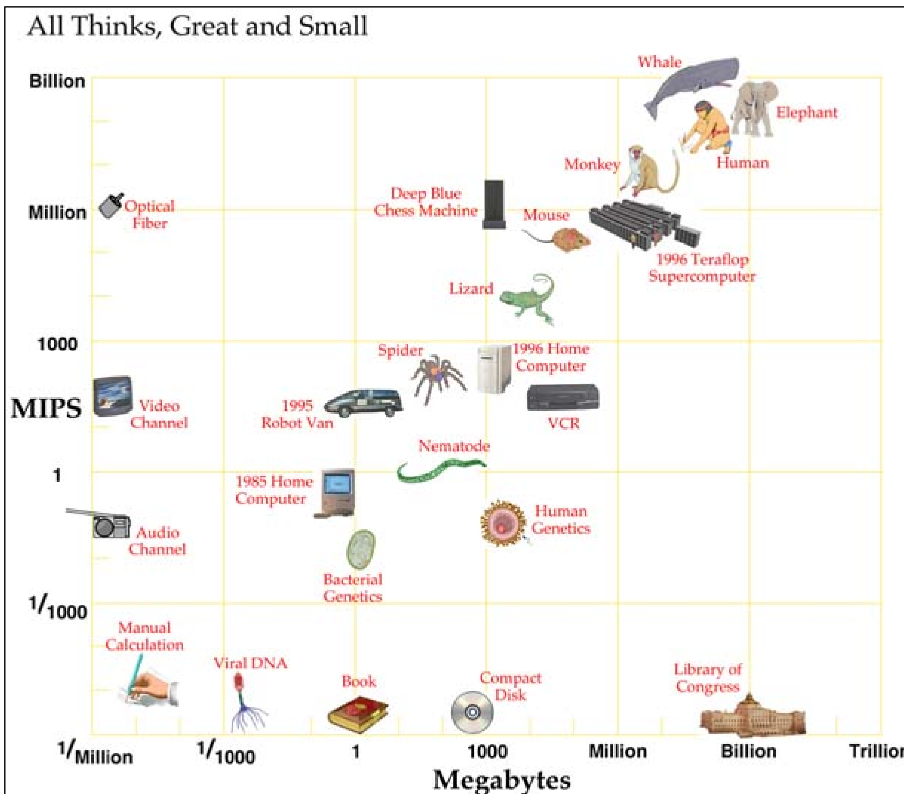
3D Memory Integration Platforms

- Novel Opportunities for Large-Scale, Heterogeneous, Hi-Bandwidth/Lo-Power Memories

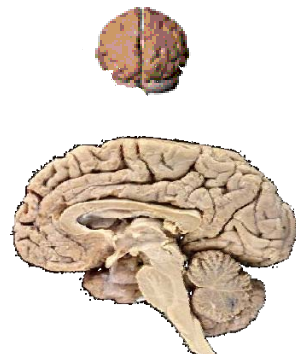
James Jian-Qiang Lu
luj@rpi.edu
Rensselaer Polytechnic Institute

SRC/NSF/A*STAR Forum on 2020 Semiconductor Memory Strategies: Processes, Devices, and Architectures
A*STAR, Singapore, October 20-21, 2009

The RPI 3D research has been funded through the Interconnect Focus Center (IFC) sponsored by DARPA, MARCO and NYSTAR, with additional support from SRC, GRC, NSF, NASA, SEMATECH, IBM, Freescale, EVGroup, Dow Chemical.



Memory of Human Brain



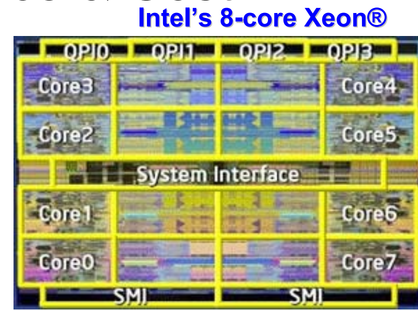
- Single type or multi-type?
- 2.5D or 3D?
- How are the "memories" and "processors" integrated?

Source: When will computer hardware match the human brain? Hans Moravec, 1998

Memories

- Memory Size, Density, Performance & Cost
- Existing Memories:
  - Registers
  - Cache, SRAM
  - Primary Memory, DRAM
  - Secondary Memory, NAND, FCRAM, SSD, Disc
  - Cross-Point (or Bar) Memory & Probe-Based Memory
- Memory/Processor ("single" chip multi-thread, multi-core processor)
- Many Challenges Ahead

Intel's 8-core Xeon®



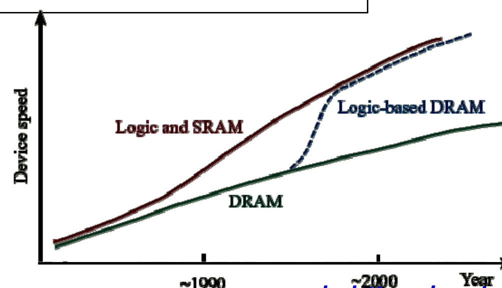
Anticipated Memory Hierarchy Capacities and Latencies in 2016

Technology	Access Time	$ per GB	MB per $
SRAM	0.5 – 5 ns	4000 – 10,000	0.1 – 0.25
DRAM	50 – 70 ns	100 – 200	5 - 10
Flash	*~ 0.2* ***ms***	*~ 60*	*17*
Magnetic Disk	5 – 20 ***ms***	0.5 – 2	500 - 2000

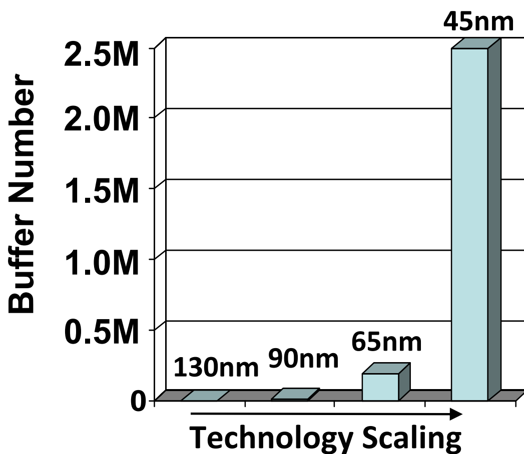
Type	Capacity	Latency
Registers	< 1 kB	~ 0.25 ns
Cache, SRAM	< 16 MB	~ 0.5 ns
Primary Memory, DRAM	< 16 GB	~ 80 ns
Secondary Memory, Disc	> 100 GB	~ 5 *ms*

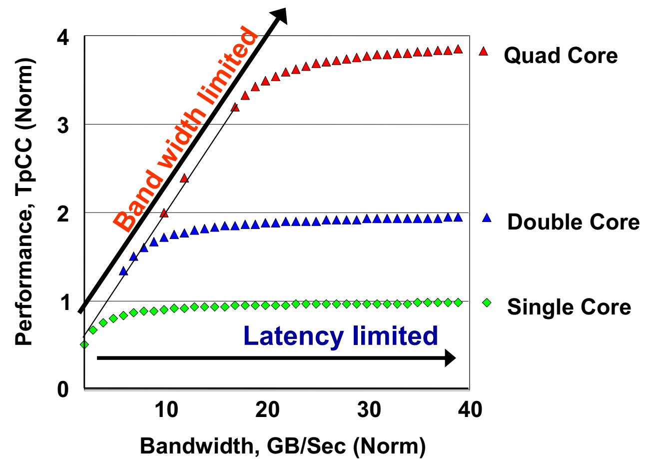
x1000 difference

- There is a difference in **capacity** and **latency** of about ***three orders of magnitude*** between the memory types
- 3D Integration for Memories?



System Integration – Technology Challenges





High resistance drives repeater challenges

Power for buffers being the leakiest and accounting for > 50% of logic leakage

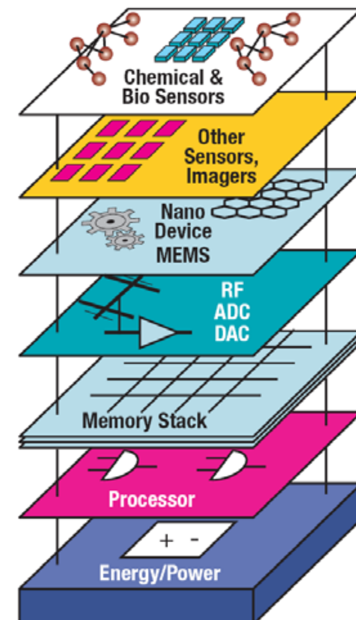
Ruchir Puri, IBM

3D extends the transfer of performance from device to the core level

Adapted from K. Bernstein, IBM

What is 3D Integration?

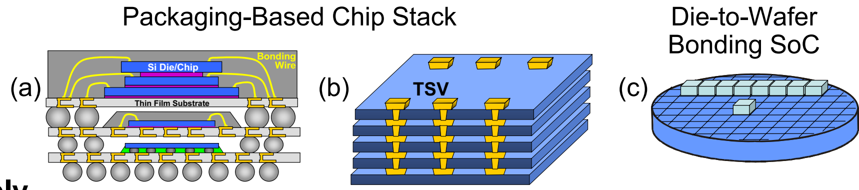
- Vertically stack and interconnect heterogeneous materials, technologies & functions, to enable extremely high functionality & bandwidth with low driving power & reduced cost
- Enhance system performance with existing/best materials, process technology, and fabrication facilities for each stratum
- Smaller, faster, cheaper and smarter

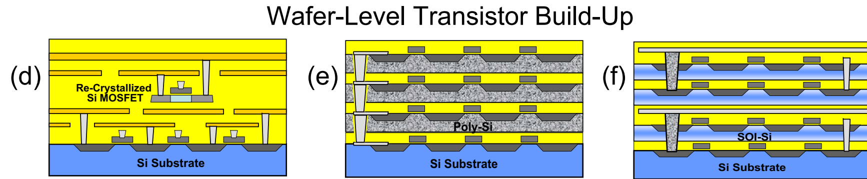


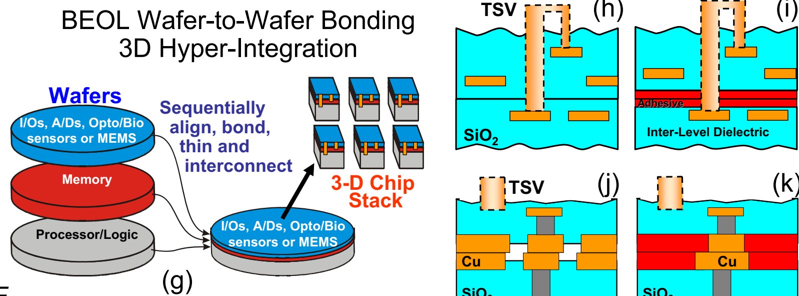
A New Paradigm for Future Technologies

3D Integration Approaches

- **3D Packaging**
- **Die-on-Wafer Assembly**
- **Transistor Build-Up**
- **Wafer-Level BEOL-Compatible 3D Hyper-Integration**

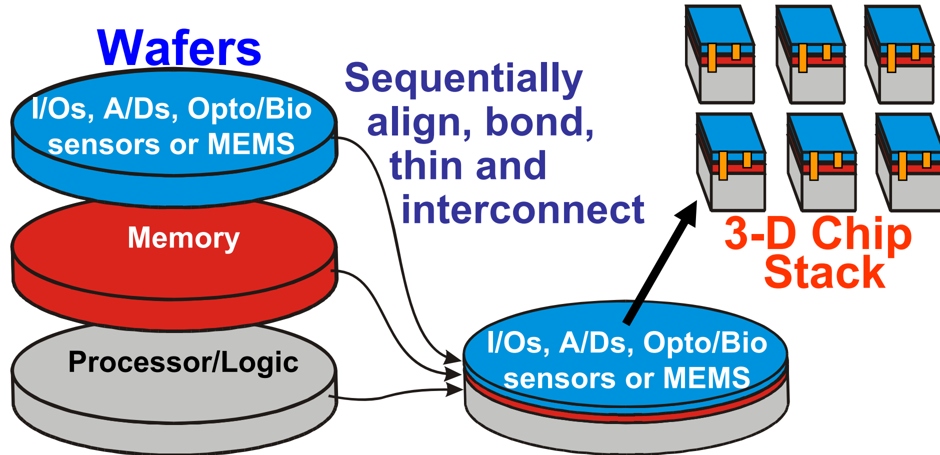






J.-Q. Lu, *Proceedings of The IEEE*, Vol. 97, No. 1, pp. 18-30, January 2009.

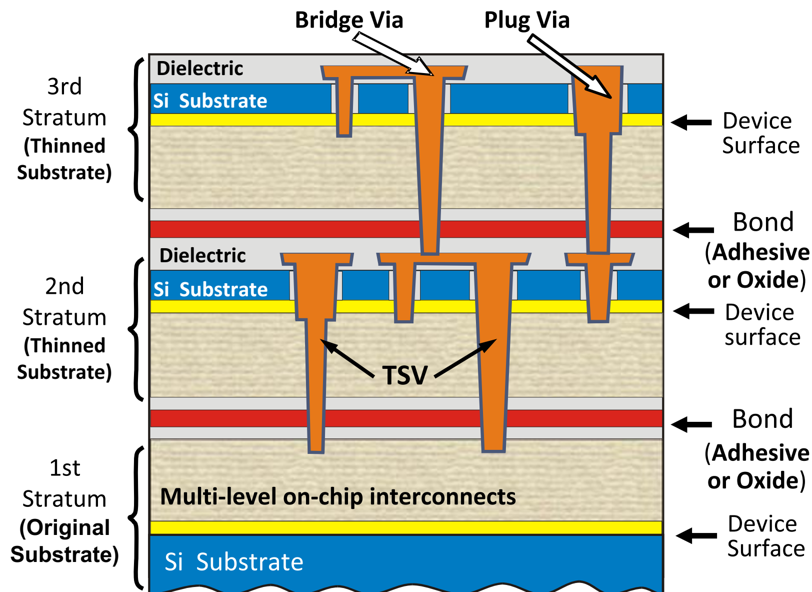
Monolithic Hyper-Integration Platform for Wafer Level 3D Integration



Increase Spead: ~ $N^{3/2}$
Reduce Power: ~$1/N^{1/2}$

- Shorten on-chip long wires (high performance)
- Small/short inter-chip via size (high density, low coupling)
- Function-specific processing (alleviating material/processing constraints)
- Monolithic process (lower cost for high-volume interconnect)

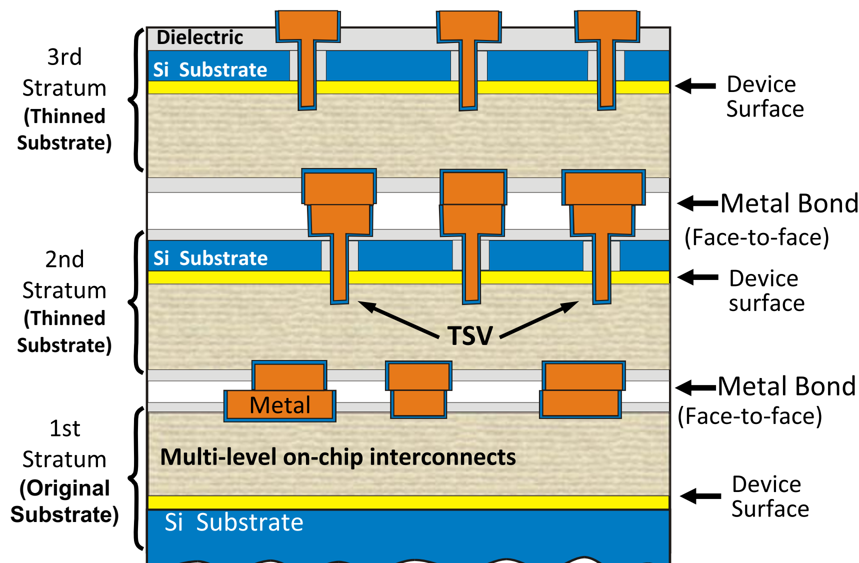
3D Platform
Based on Dielectric Bonding
(Post-Bonding TSVs)



- Separate, robust wafer bonding
- TSVs pass through thinned Si, FEOL & BEOL layers to connect ICs on two strata

- Face-to-Face without Handle Wafer
- Back-to-Face with Handle Wafer

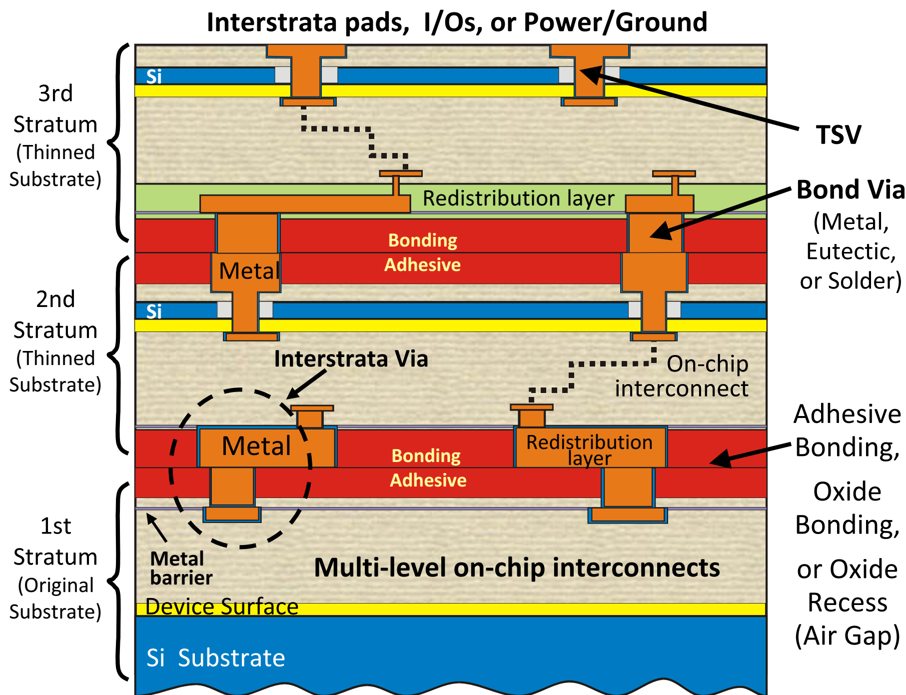
3D Platform
Based on Metal Bonding
(Prior- or Post-Bonding TSVs)



- One-step to bond wafers and connect ICs on two strata
- TSVs pass through thinned Si & FEOL layers

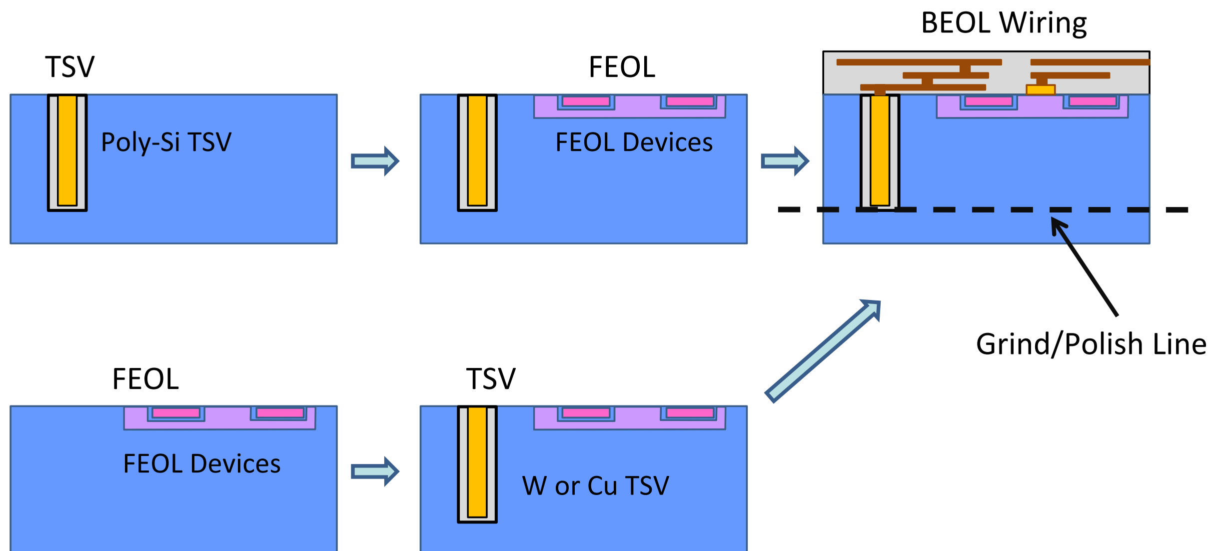
- Face-to-Face without Handle Wafer
- Back-to-Face with Handle Wafer

3D Platform
Based on Hybrid Metal/Dielectric Bonding
(Prior- or Post-Bonding TSVs)

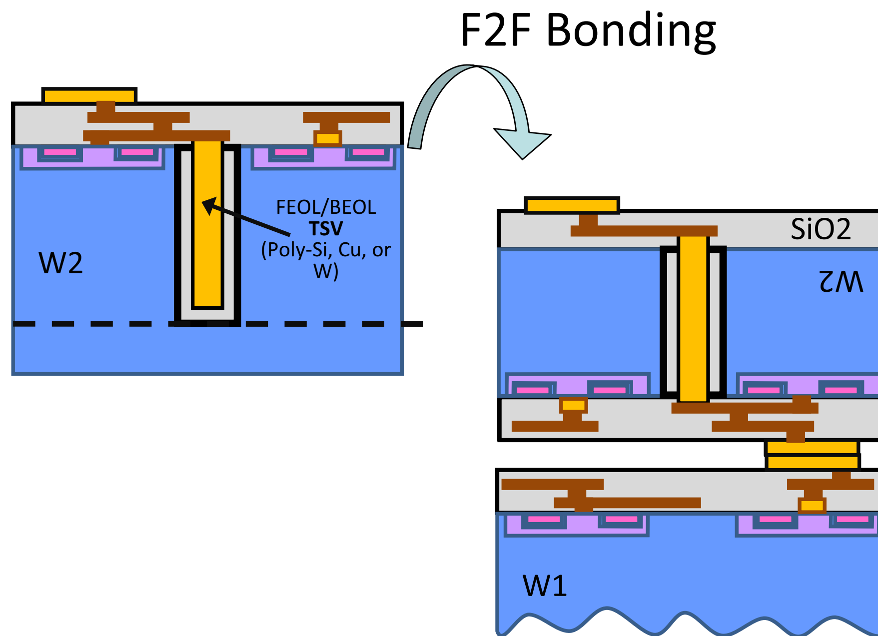


- Combine dielectric bonding and metal bonding
- Redistribution layer to alleviate routing and alignment
- TSVs pass through thinned Si & FEOL layers

FEOL and BEOL TSV Formation



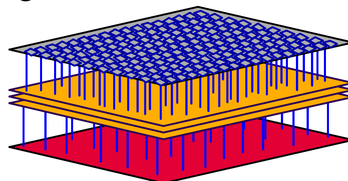
Process Sequence w/o Handle Wafer



luj@rpi.edu

Wafer-Level 3D IC: Benefits, Challenges and Strategies

Potential Benefits
- Density/Capacity
- Wire Delay
- Bandwidth
- Power
- Material/Processing
- Functionality
- Packaging
- Cost



Challenges
- Thermal and Power
- Yield concerns
- Test methodology
- Infrastructures
  - Equipment
  - ECAD Design
  - Standard

Technology Strategies
- Platform/Architecture
  - Face-to-face vs face-to-back
  - Via-first vs via-last
  - Transistor build up vs. BEOL 3D
- Substrate
  - SOI (Si-on-Insulator)
  - Bulk Si
- Processing Technologies
  - Alignment
  - Bonding
  - Thinning
  - Inter-wafer interconnect

- **Eliminate the slower and higher-power off-chip buses to memory by replacing them with high-bandwidth and low-latency vertical interconnections**

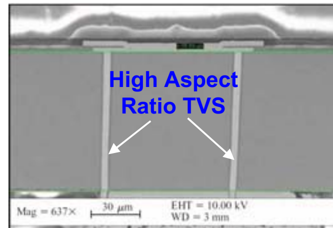
luj@rpi.edu

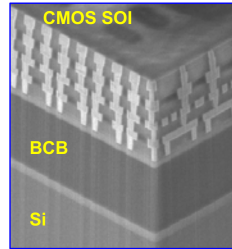
3D Integration Process Advances

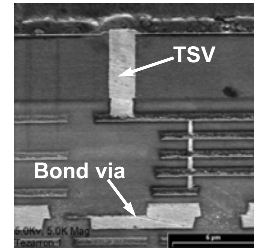
- Alignment: ~ 1 μm
- Bonding: ≤ 400 °C
- Thinning: easy for SOI, difficult for bulk Si
- Inter-Wafer Interconnect:
  - Through-Si-Via (TSV) or Through-Strata-Via
  - Bond via, formed at bonding interface

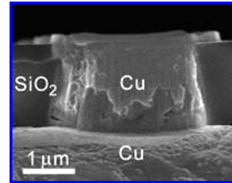
TSV pitch: > 5 μm

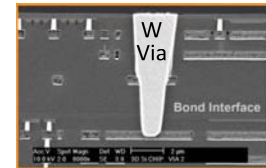
Note: Cross-section images are adapted from papers published by RPI, IBM, MIT LL, Fraunhofer, Tezzaron

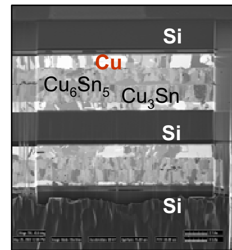


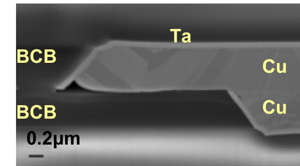


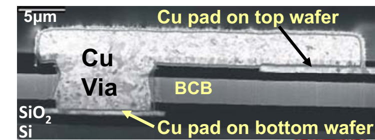










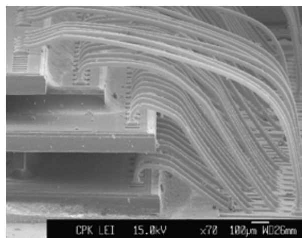


TSV 3D IC Technologies & Equipment

Chip or Wafer bonding		Via drilling	Via filling	Thinning
C2W	W2W			
Technologies				
Adhesive, fusion oxide or metal-metal bonding Chip alignment	Adhesive, fusion oxide or metal-metal bonding Wafer alignment	Laser drilling DRIE Photolithography	Electroplating CVD Photolithography	Grinding (BG) CMP Wet etching Plasma etching
Equipment				
Flip Chip Bonder	Wafer bonder Wafer aligner	Laser OR DRIE equipment Coater Mask aligner OR stepper	Metal deposition system Coater Mask aligner OR stepper	Thinning equipment Temporary bonding equipment

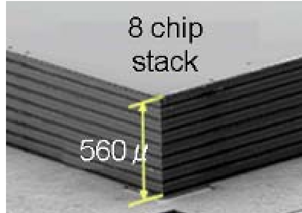
Memory Stacks

Wire-bonded

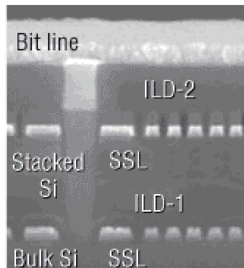


ChipPAC Inc.

TSV-Connected NANDs

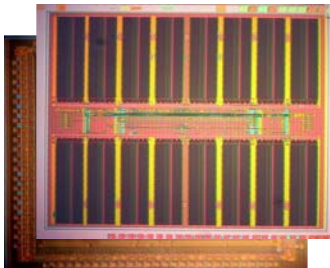


Transistor Build-up



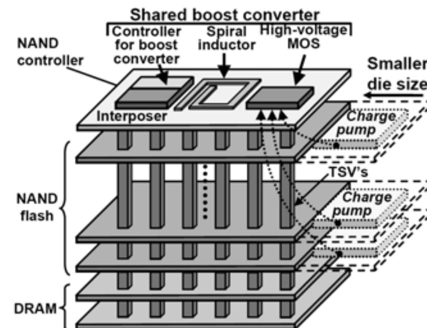
Samsung

Wafer-Level 3D DRAMs



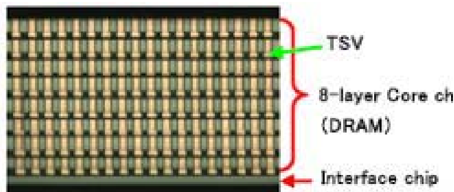
Tezzaron

3D SSD w/ Shared Boost Converter



Toshiba, IEEE 3DIC 2009

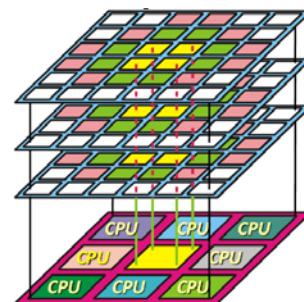
TSV-connected 8 Gb DRAM



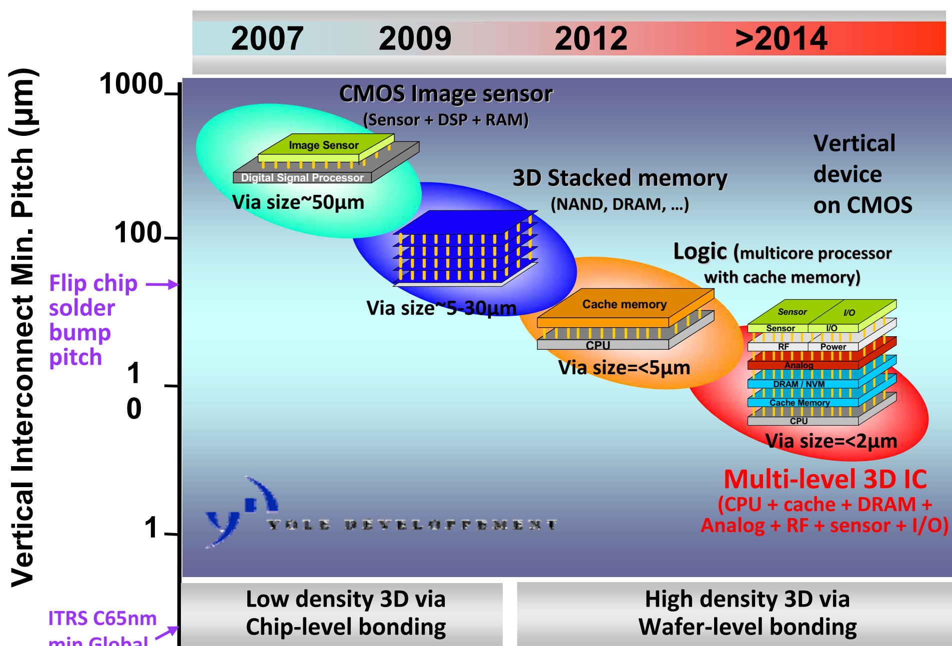
Elpida

- **Drastic power reduction for driving I/Os**

Memory/Multi-Core



3D Market Timing Projections



http://www.yole.fr/

TSV CoO

- "The EMC^{3D} consortium, which has been working on an interconnect Thru-Silicon-Via, or TSV process flow, announced that the cost of ownership (CoO) for a 10,000-wafers-per-month capacity line is now achievable at less than $190 per wafer."
  - – September 4, 2008
- Updated Goal for 2009 under consideration: < $150.00/wafer

http://www.emc3d.org/

Suggestion of Research Areas

- Basic research: materials, processing, reliability, thermal, modeling & design CAD tools
- Simulation and compact models for prediction of density, speed and energy of 3D memory stack
  - Impact of TSV electric/magnetic/thermal performance and placement/routing?
- Impact of 3D integrated memories on µProcessor performance
  - CPI beyond the memory wall?
  - Do we need Network on Chip (NoC) for many core 3D systems with huge stacked memory?
- Algorithm & hierarchy for NEW 3D memory architectures
  - Improve density, speed and energy, e.g., with separate strata of addressing, driving, or sensing circuitry?
  - Allocate shared memory resources between multiple cores?

Perspectives

- 3D integration technology with many advantages is almost ready for its prime time; 3D era has just started
  - Feasibility is demonstrated, though many questions are to be answered;
  - Detailed basic research (materials, processing, reliability, thermal, modeling & design CAD tools) and R&D evolution are under way.
- 3D technology enables heterogeneous memory integration with high bandwidth, low latency, low driving power and simplified process (high yield/low cost)
- Known-Good-Die (KGD) has been answered with redundancy, ECC, and simplified process
  - DFT, BIST, ECC and redundancy (**memory & processors**) will be important to 3D yield, test & cost
- 3D integration will facilitate NEW approaches to memory/processor design and manufacture
- 3D integration may enhance IC security and yield new software design, such as for virtual machine (software stack)