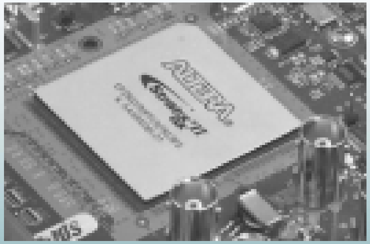
3D Super-Via for Memory Applications

Hong Sangki (shong@tezzaron.com)

Tezzaron Semiconductor Corporation

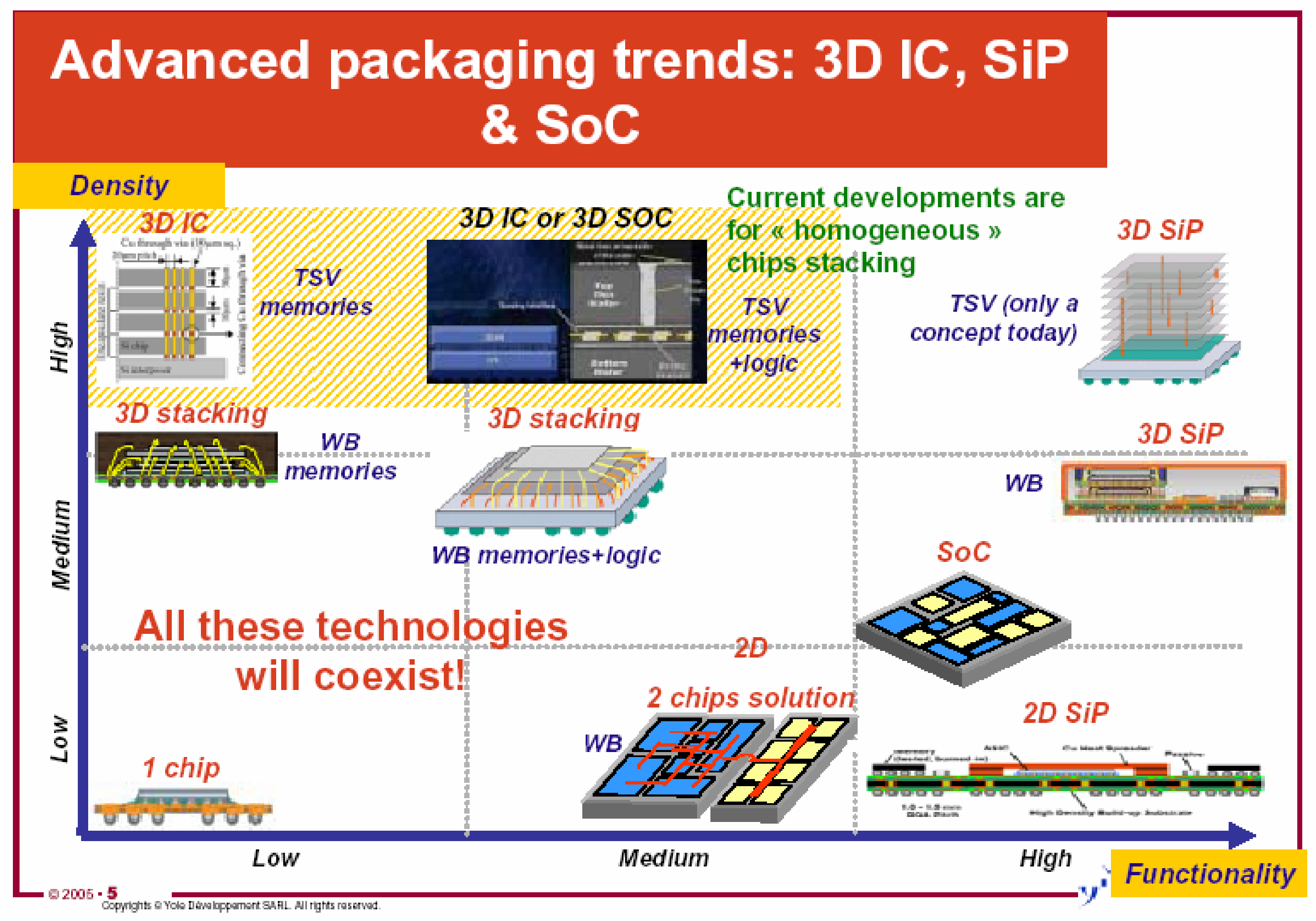
Chicago - Singapore

Market Drivers for 3D

Density/ Functionality	Mobile & Wireless	
Performance ✓ Faster ✓ Low Power	Workstations Super-computer	
Cost/ Yield	Scaling cost Yield Packaging Yield	

Market Drivers for 3D Memory

Driver	Functionality	Technical Parameter # 1	Technical Parameter # 2	Value Indicator
Stacked NAND Flash	Cell Phones Hard Drives Flash Drives	Memory density		High packing density
Microprocessor + Memory	Workstations	Latency bandwidth	Power	Execution time
Memory	Multiple	Density	Latency	Varies



Advanced packaging trends: 3D IC, SiP & SoC

Density

High

Medium

Low

3D IC

TSV memories

3D IC or 3D SOC

TSV memories +logic

Current developments are for « homogeneous » chips stacking

3D SiP

TSV (only a concept today)

3D stacking

WB memories

3D stacking

WB memories+logic

3D SiP

WB

SoC

All these technologies will coexist!

2D

2 chips solution

WB

2D SiP

1 chip

Low

Medium

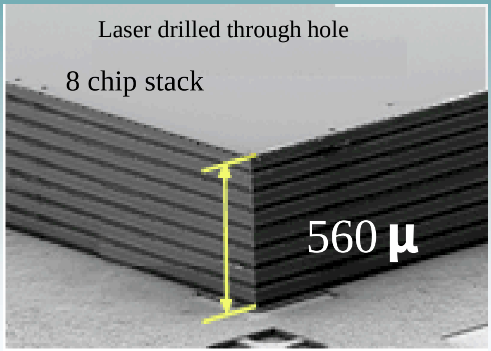
High

Functionality

© 2005 • 5

Courtesy of Yole Développement

How Real is 3DMemory?

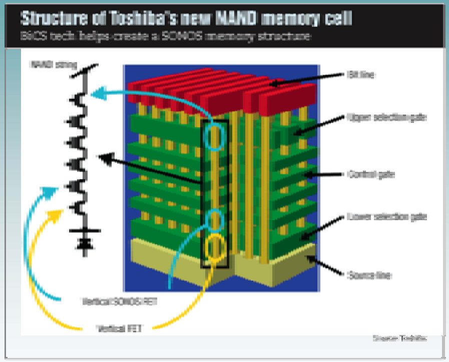


Samsung

16Gb NAND flash

(2Gx8 chips),

560 µ thin

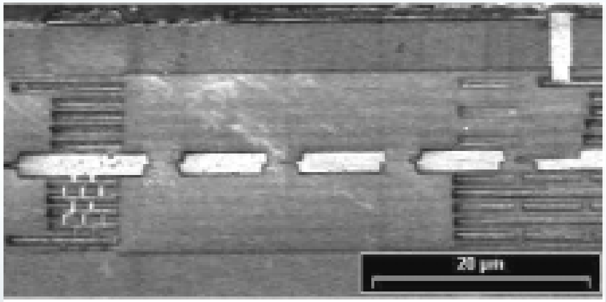

Structure of Toshiba's new NAND memory cell

BiCS tech helps create a SONOS memory structure

Source: Toshiba

Toshiba

3D NAND



Tezzaron

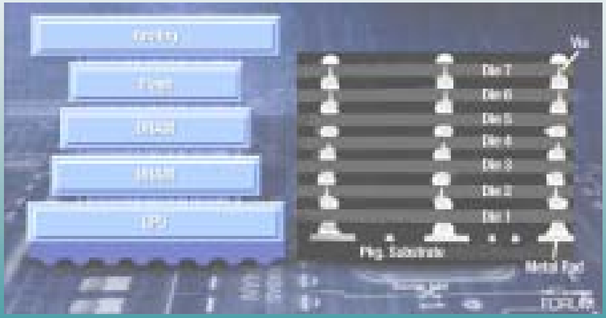
CPU + memory

Micron

Osmium Memory Stacking

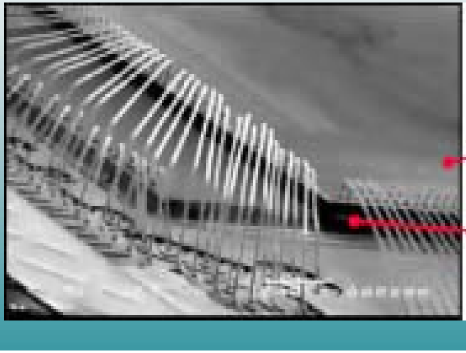
NEDO

1Gbit DRAM (128Mbx8 chips) 5mm2



Intel

CPU + memory



TI

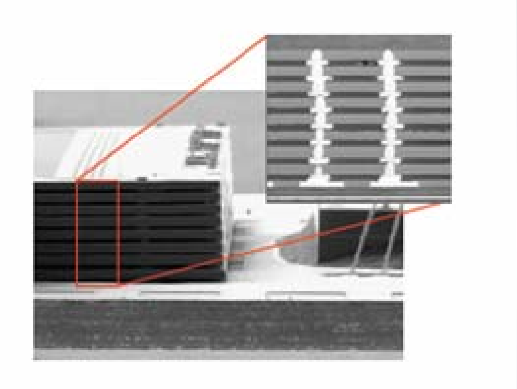
OMAP1611 Applications Processor

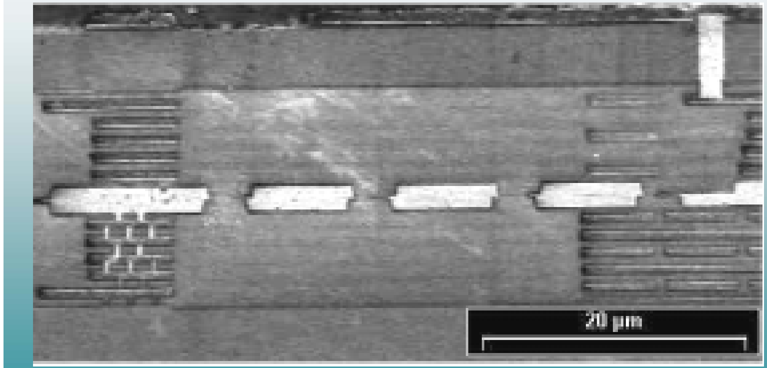
256 Mbit Mobile DDR SDRAM

The 2007 ITRS Roadmap will include 3D!

- Tables for TSV size, pitch
- Bond point pitch
- Wafer thickness

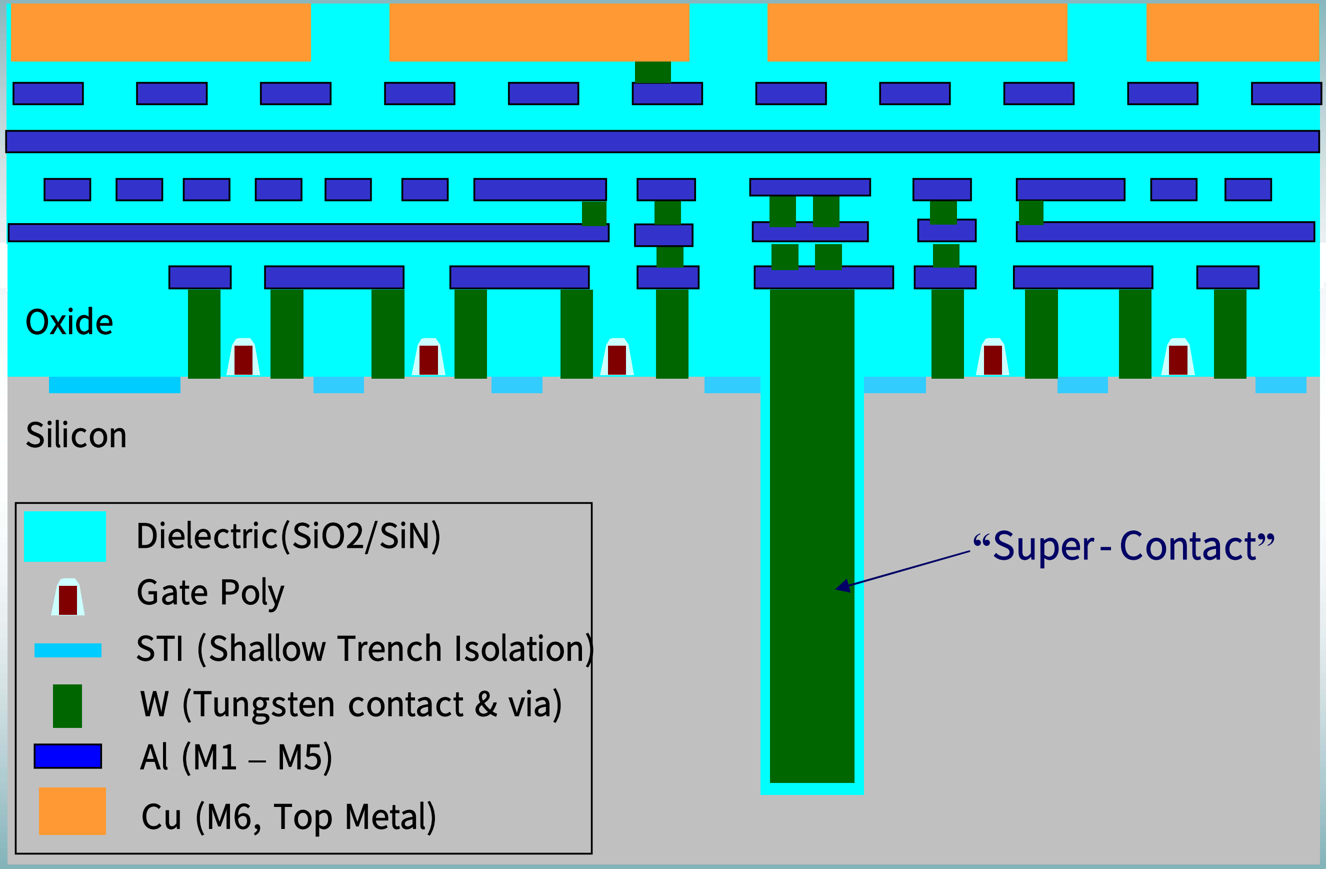
Low density:
$10^2 - 10^4$ (> 50 µm pitch)
Wafer Scale Packaging





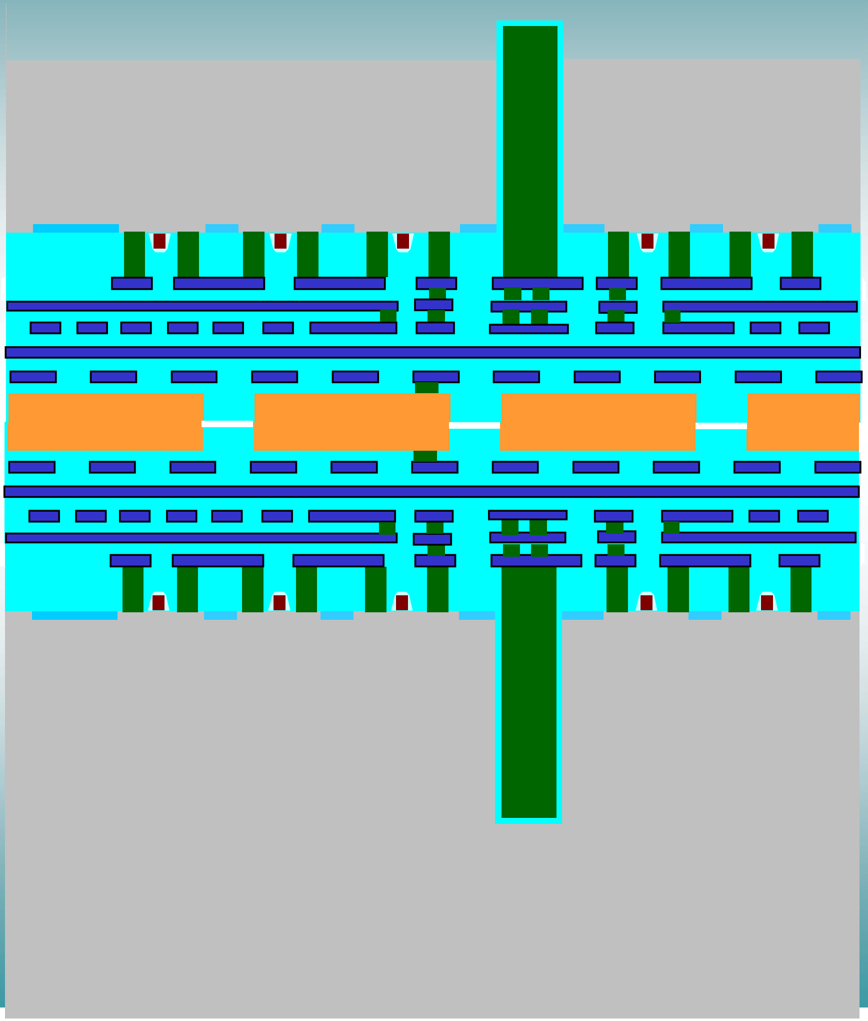
High density:
$10^5 - 10^6$ (5 ~ 25 µm pitch)
Wafer Scale Package

A Closer Look at Tezzaron's Stacking



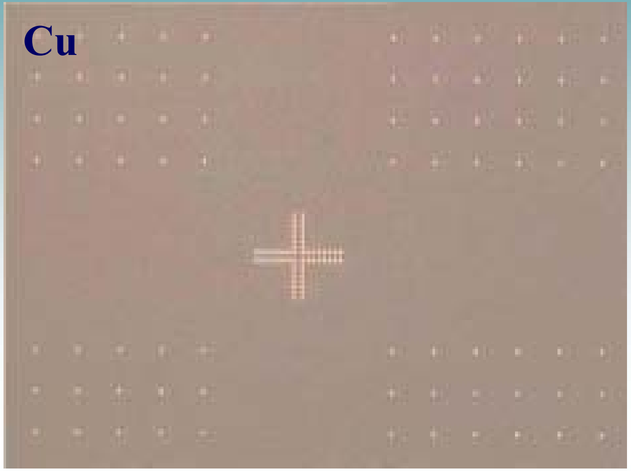
Courtesy, Tezzaron Semi.

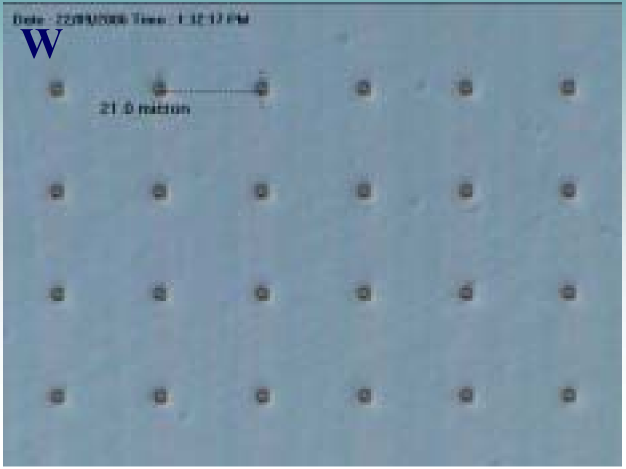
2nd Wafer Stack & Thin

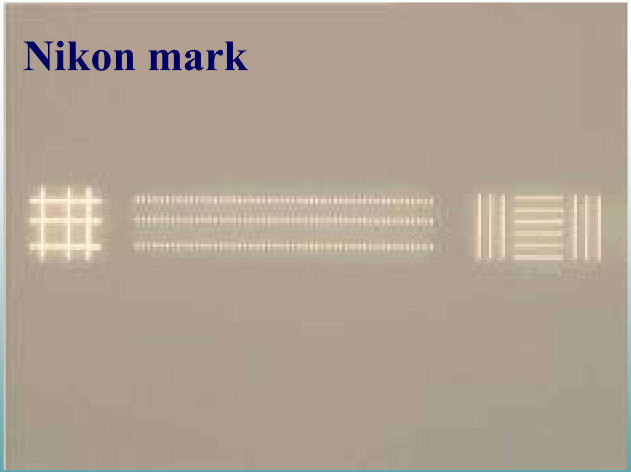


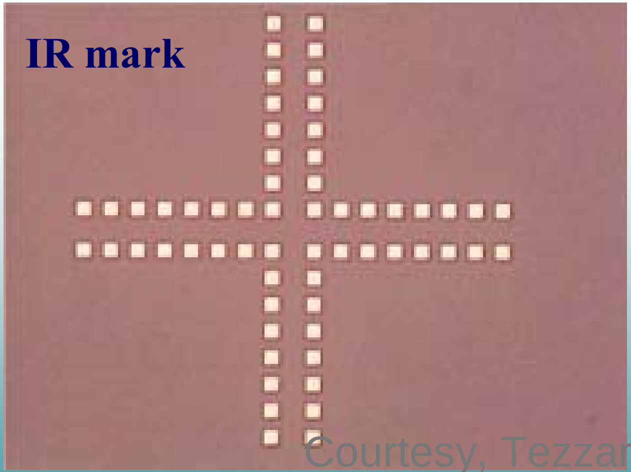
Courtesy, Tezzaron Semi.

Wafer Backside after thinning



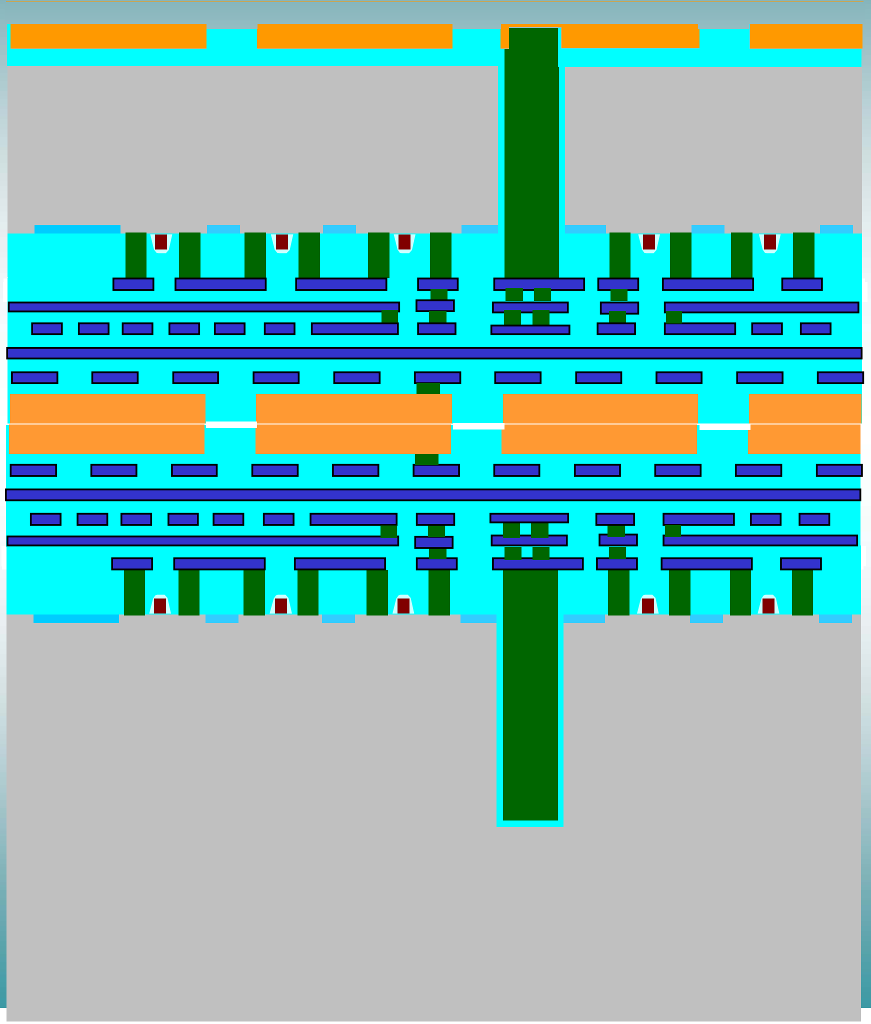






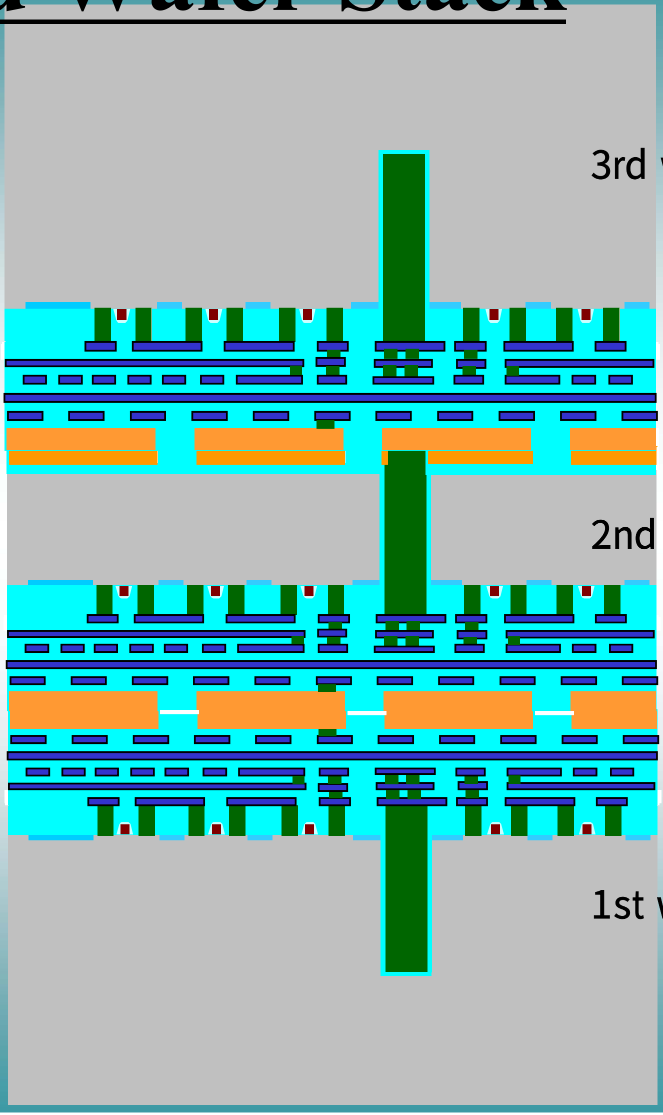
Courtesy, Tezzaron Semi.

Backside Cu Metallization



Courtesy, Tezzaron Semi.

3rd Wafer Stack



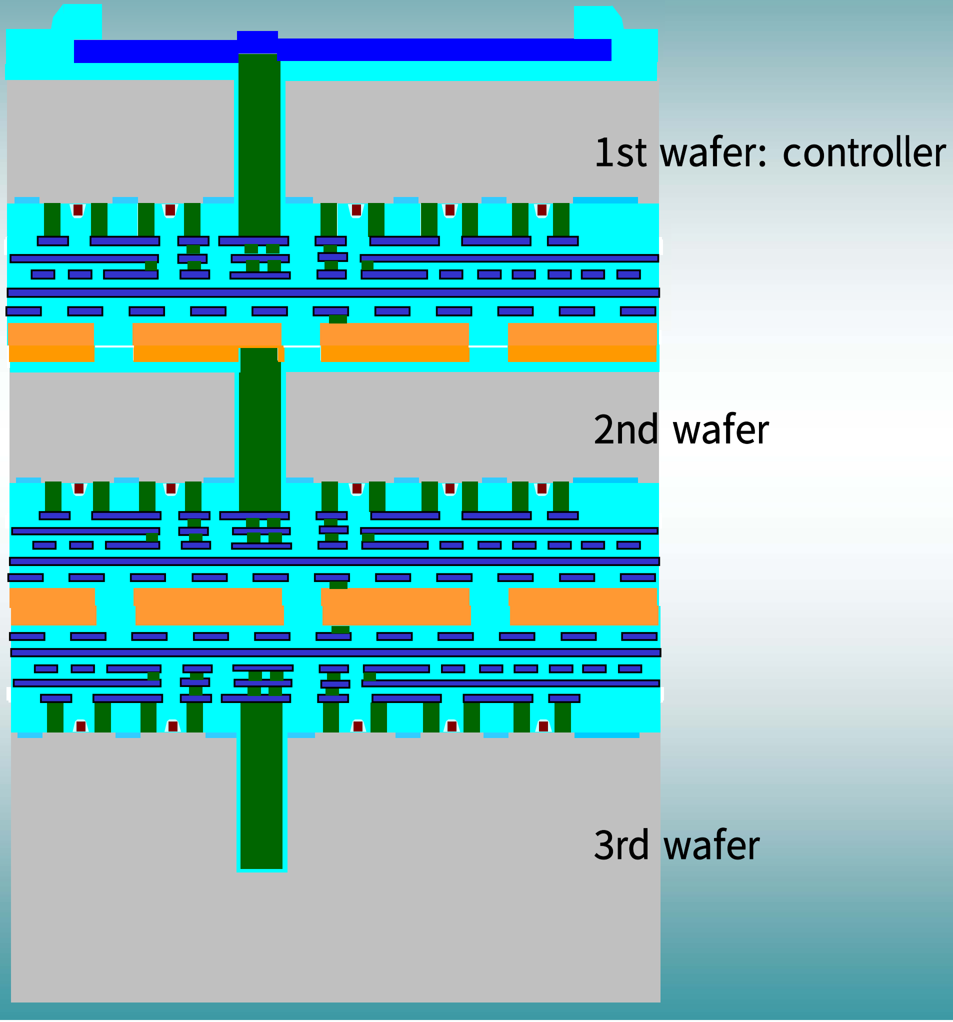
3rd wafer

2nd wafer

1st wafer: controller

Courtesy, Tezzaron Semi.

Flip & I/O Pad Out



Courtesy, Tezzaron Semi.

Stacking Process Sequential Picture

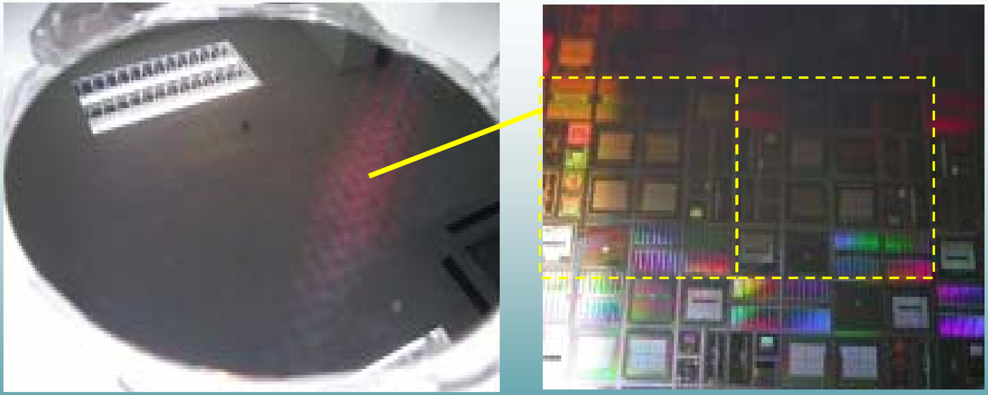
Two wafer Align & Bond

Course Grinded

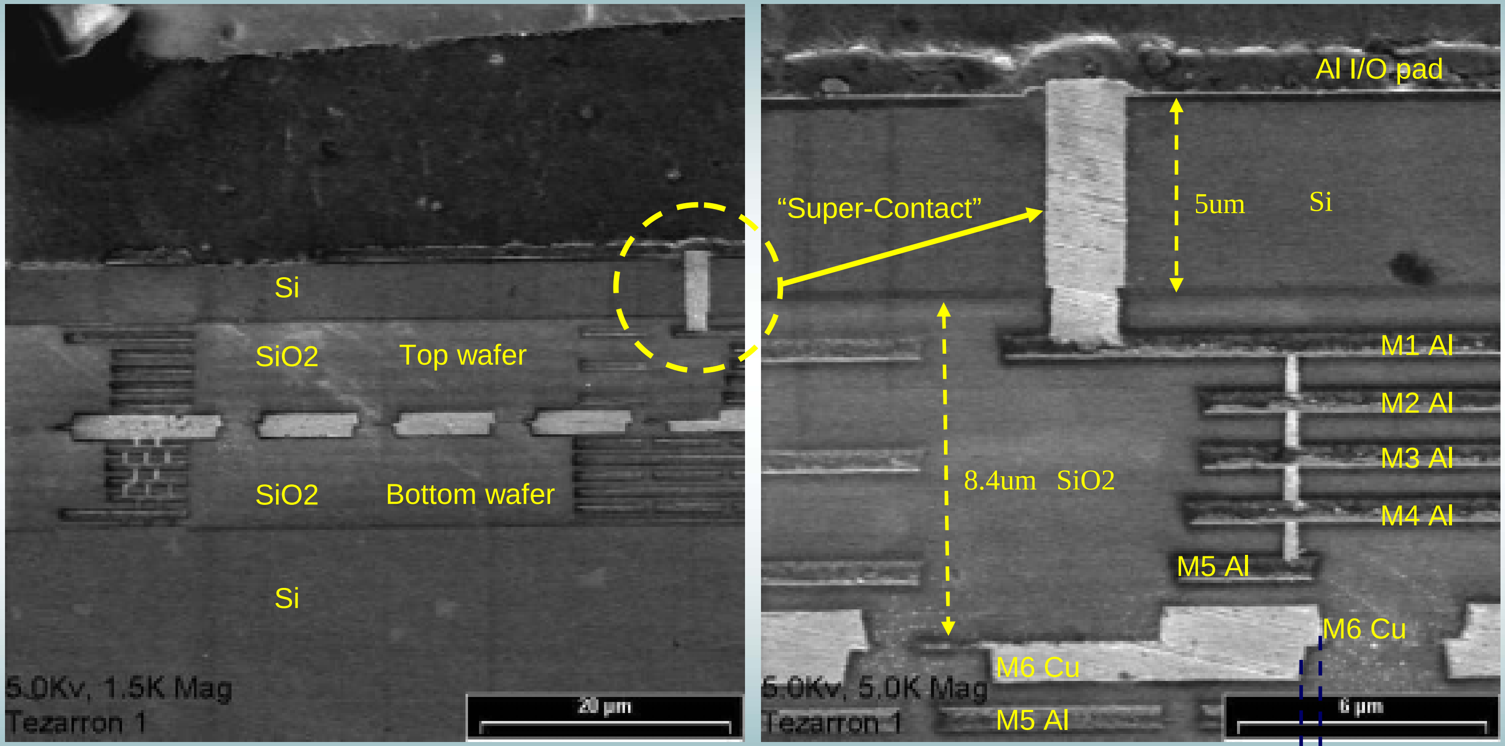
Fine Grinded

After CMP

Si Recessed



CPU + SRAM cross-section

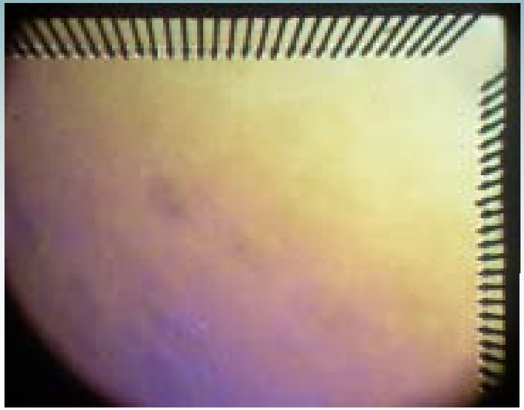


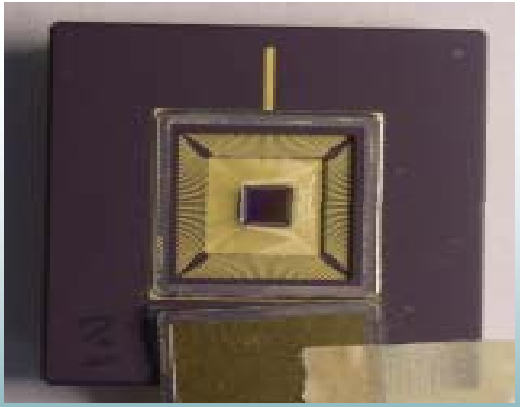
Courtesy, Tezzaron Semi.

Wafer-to-wafer misalign ~0.4um

CPU/Memory Stack

- R8051 CPU
  - 80MHz operation; 140MHz Lab test (VDD High)
  - 220MHz Memory interface
- IEEE 754 Floating point coprocessor
- 32 bit Integer coprocessor
- 2 UARTs, Int. Cont., 3 Timers, …
- Crypto functions
- 128KBytes/layer main memory
- Completely synthesized, placed and routed in 3D with standard Cadence tools. Runs slightly better than predicted by models and tools.



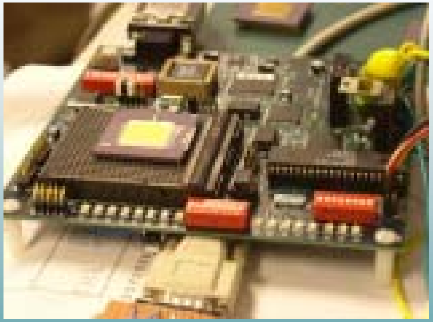


Courtesy, Tezzaron Semi.

Speed gaining demo by Stacking Memory on CPU

Tezzaron's Stacked Memory on CPU

MCM package of Memory and CPU



Courtesy, Tezzaron Semi.

Results & Demonstration of 3D CPU

Stacked 8051 CPU vs. Dallas semiconductor 80C420

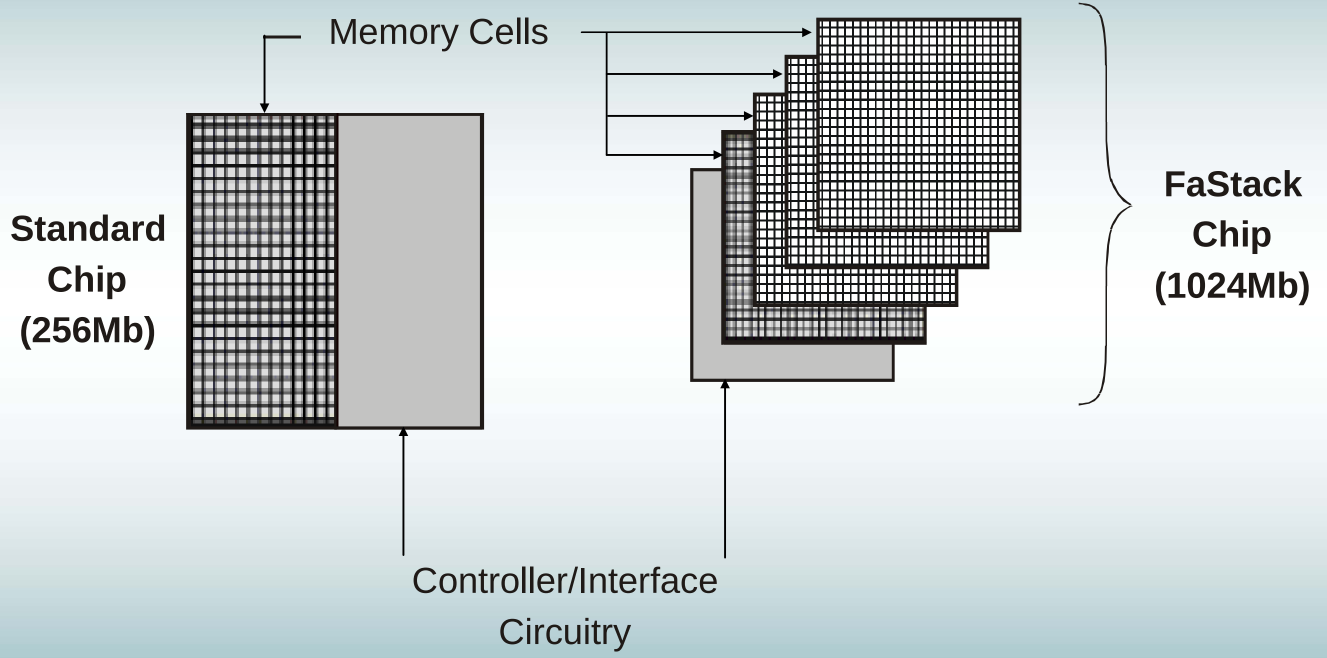
Video (split half screen)	*3X faster*
Complex Math Calculation	*4.5X faster*
Memory Operation	*7X faster*
Current consumption	*3X lesser*
Power Consumption	*10X lesser*

Courtesy, Tezzaron Semi.

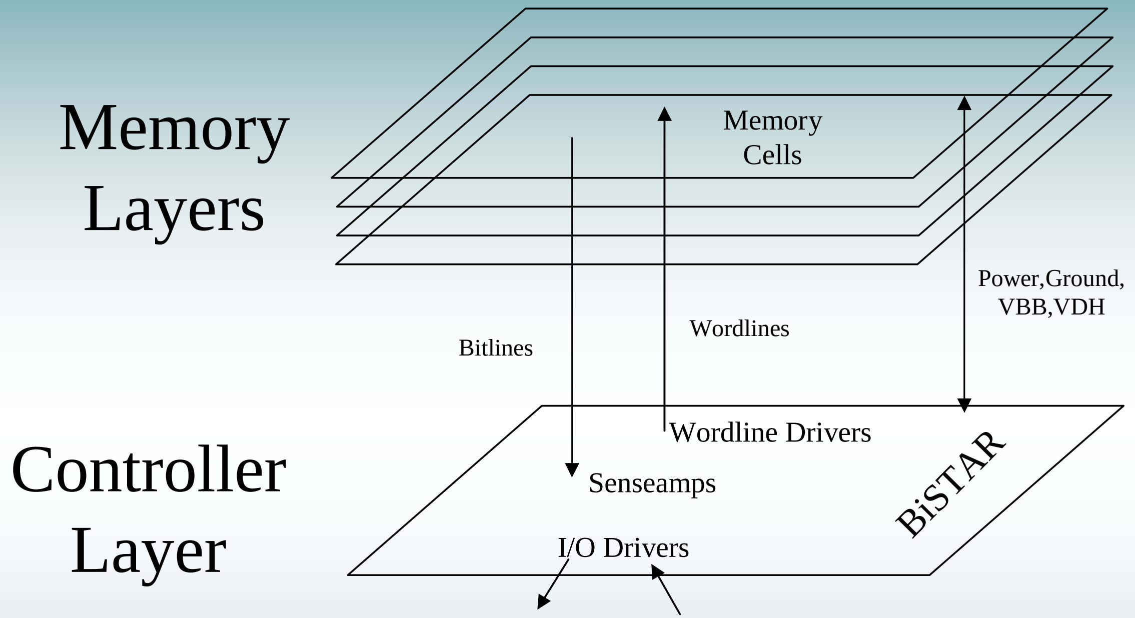
Results of early Qual data

- 100,000 device thermal cycles (–65 to 150C 15 minute soak)
  - No failures
  - Two build lots
- 168 hour high temp
  - No Failures
  - Extended to 336 and then 504 with no failures
- Hot spot delamination testing
  - >10watts/sqmm, no failures
- Life test under bias
  - >10,000 hours, no failure

Concept of Tezzaron's 3D DRAM



“Dis-Integrated” 3D Memory

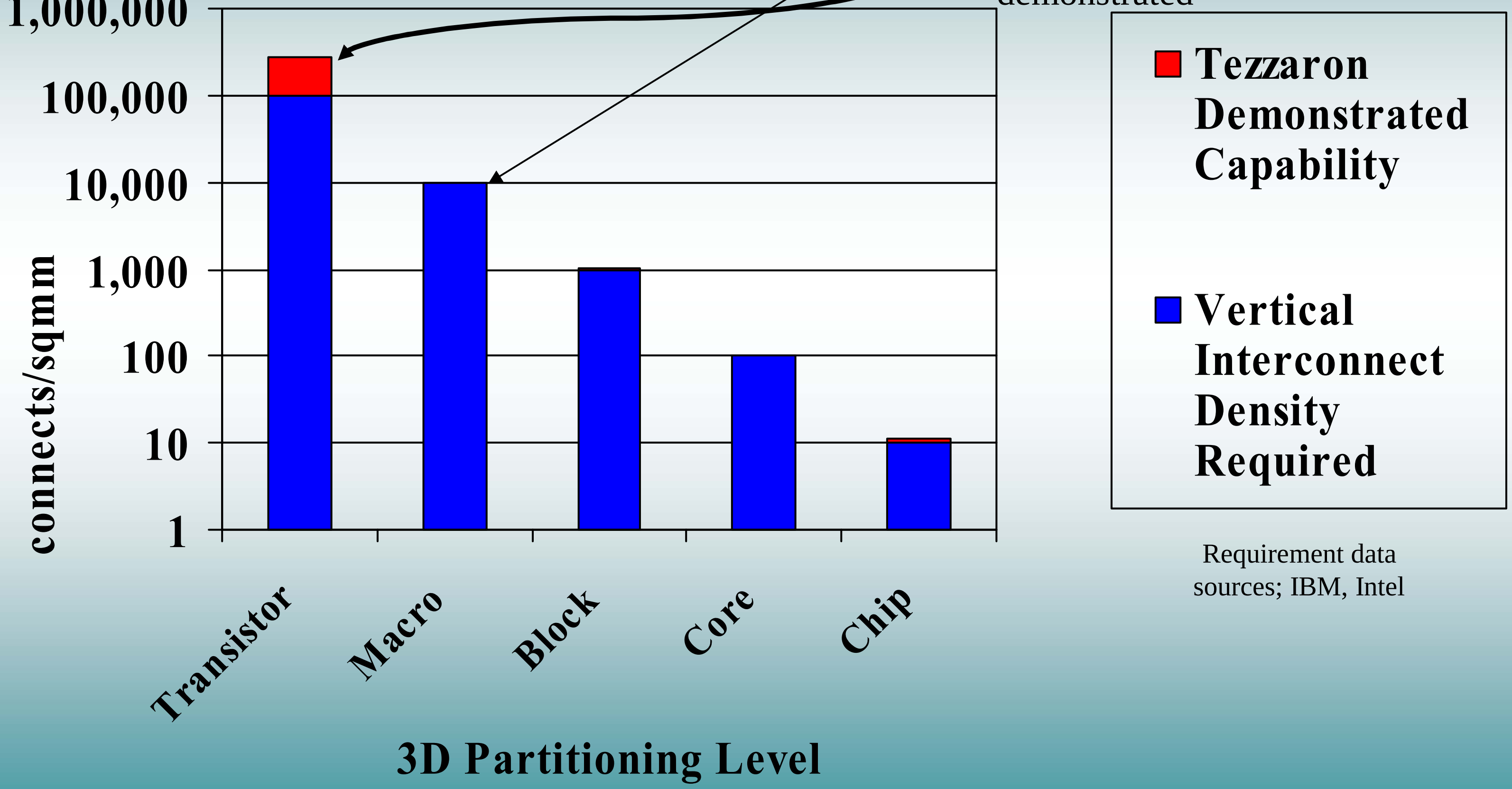


But…..this requires,

Millions of *vertical* interconnect!

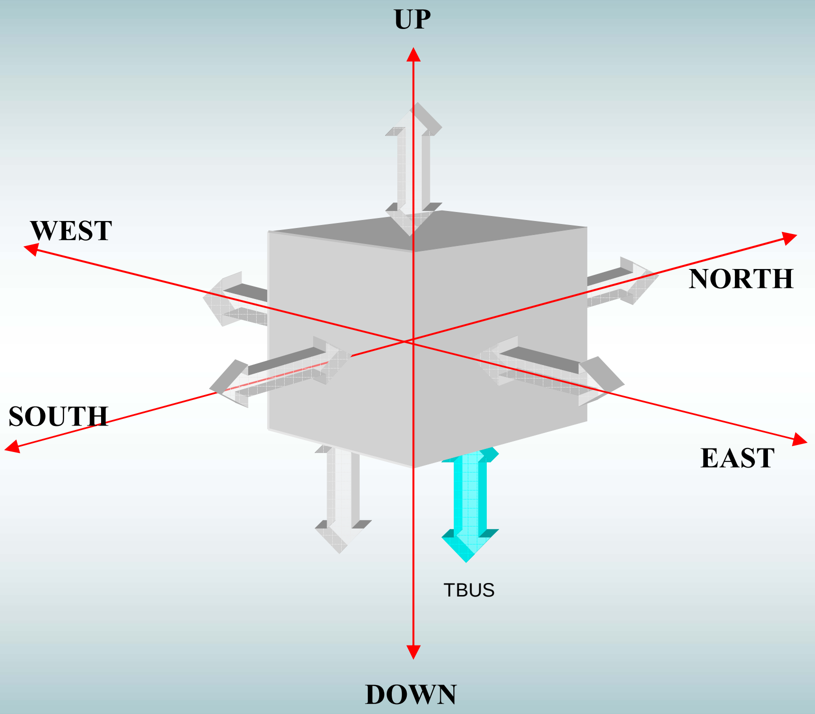
How many interconnects are required?

Most Tezzaron designs use ~10,000/sqmm, but up to 170,000/sqmm have been demonstrated

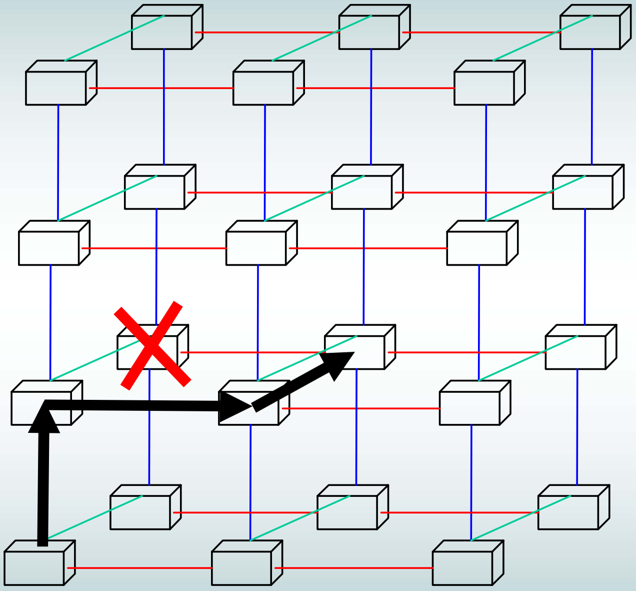


Requirement data sources; IBM, Intel

3D Interconnect



3D-Routing Node



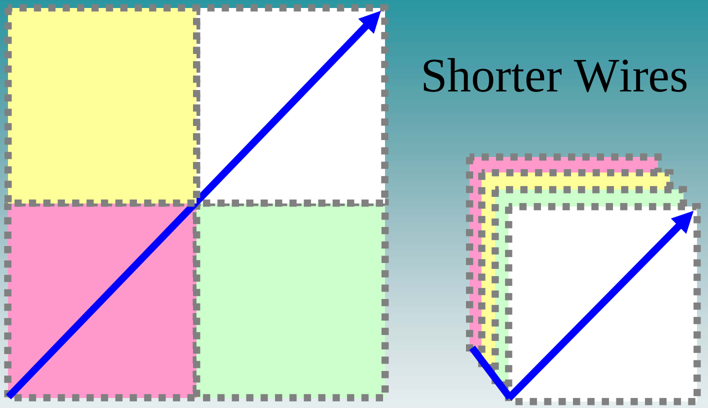
What can Tezzaron 3D DRAM Achieve?

- Faster Access Time
- Lower Power
- Denser
- Reliable
- Compatible
- Lower Costs

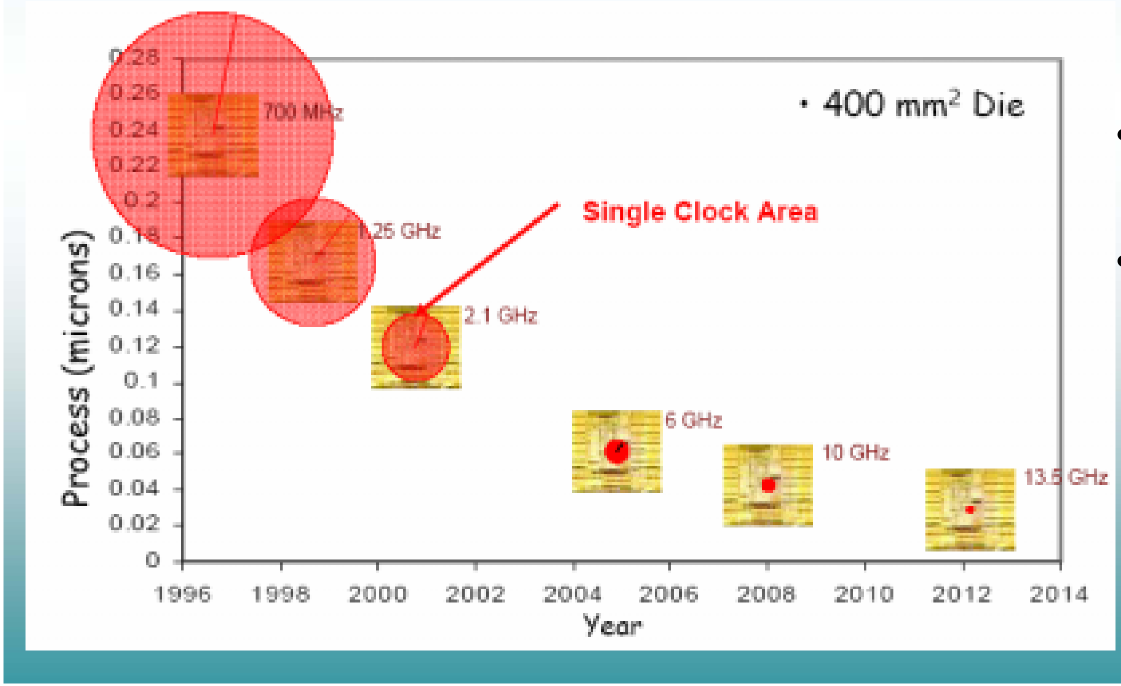
1. Faster!

Propagation delay is proportional to: $\frac{1}{\#\text{ of layers}}$

$$t_d \approx 0.35 \text{ x } rcl^2$$



Shorter Wires



- Global Interconnect "problem"
- Span of Control

… in the older 1.0 µm Al/SiO_2 technology generation the transistor delay was ~20 ps and the RC delay of a 1 mm line was ~ 1.0 ps, while in a projected 35nm Cu/low κ technology generation the transistor delay will be ~1.0 ps, and the RC delay of a 1 mm line will be ~250 ps.[i]

In addition, in the 0.13um technology node approximately 51% of microprocessor power was consumed by interconnect, with a projection that without changes in design philosophy, in the next 5 years up to 80% of microprocessor power will be consumed by interconnect.[ii]

[i] J. Davis and J. Meindl, Interconnect Technology and Design for Gigascale Integration, Kluwer Academic Publishers, 2003.

[ii] N. Magen, A. Kolodny, U. Weiser, N. Shamir, “Interconnect-Power Dissipation in a Microprocessor,” ACM System-Level Interconnect Prediction Workshop, Feb 2004

1. Faster!

DRAM wants 2 different processes!

Bit cells	Low leakage -slow refresh -low power -low GIDL	High Vt Devices Vneg Well Thick Oxide
Sense Amps Word line drivers Device I/O	High speed -better sensitivity -better bandwidth -lower voltage	Low Vt Devices Copper interconnect Thin Oxides

2. Lower Power!

$$P_{avg} = VDD \text{ x } I_{avg} = C_{tot} \text{ x } VDD^2 \text{ x } f_{clk}$$

C is mostly due to wiring

Therefore:

$$P_{avg} \propto l_{avg}$$

Or:

$$P_{avg\ stacked} \approx \frac{P_{avg\ single\ layer}}{\text{\# of layers}}$$

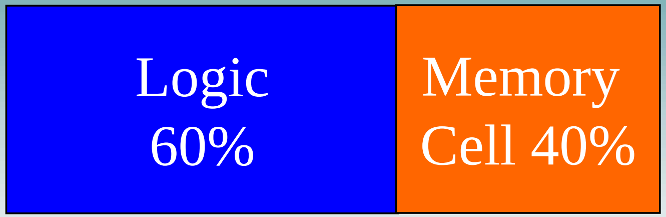
Operation	Energy
32-bit ALU operation	5 pJ
32-bit register read	10 pJ
Read 32 bits from 8K RAM	50 pJ
Move 32 bits across 10mm chip	100 pJ
Move 32 bits off chip	1300 to 1900 pJ

Calculations using a 130nm process operating at a core voltage of 1.2V (Source: Bill Dally, Stanford)

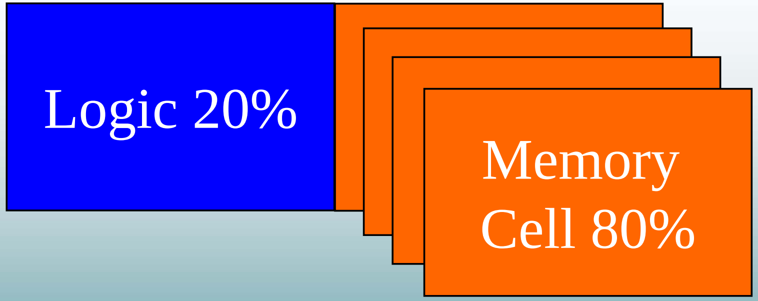
3. Lower Costs & Higher Yield!

- Less processing per layer
- Better optimization per wafer
- Higher bit density in memories
- Lower test cost using Bi-STAR™
- Higher yield using Bi-STAR™

Standard DRAM Utilization



66% Savings in logic per memory cell



Increasing Die Overhead

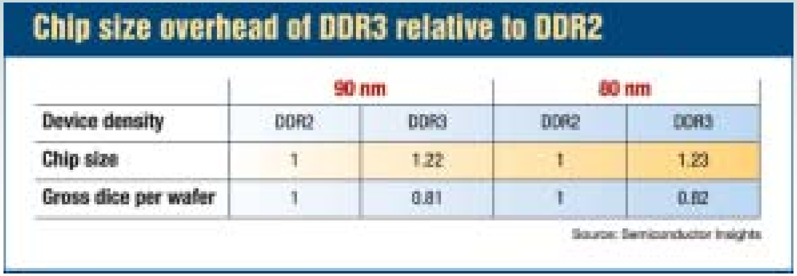
Array Utilization

DDRI 70%

DDRII 47%

DDRIII 38%

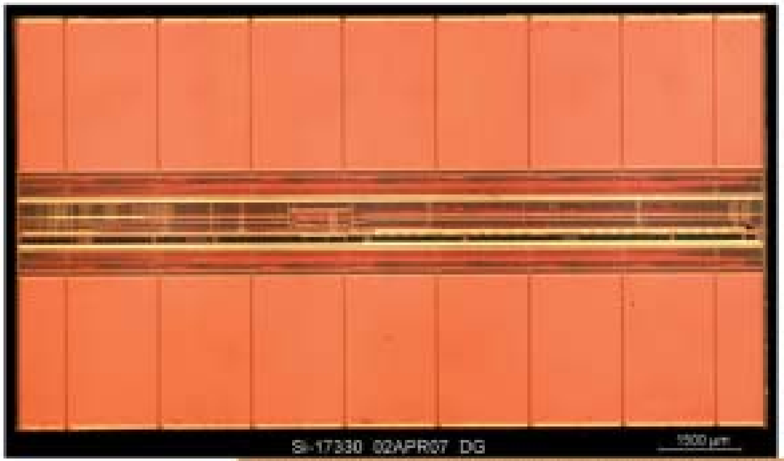
DDRIV <30%?



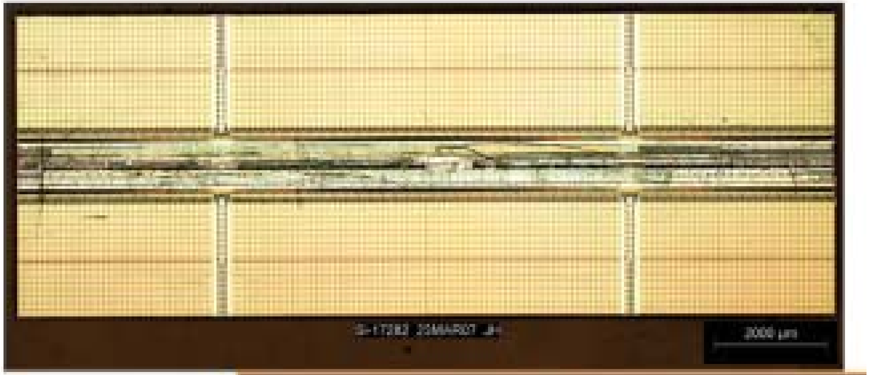
Chip size overhead of DDR3 relative to DDR2

	90 nm		80 nm	
Device density	DDR2	DDR3	DDR2	DDR3
Chip size	1	1.22	1	1.23
Gross dice per wafer	1	0.81	1	0.82

Source: Semiconductor Insights



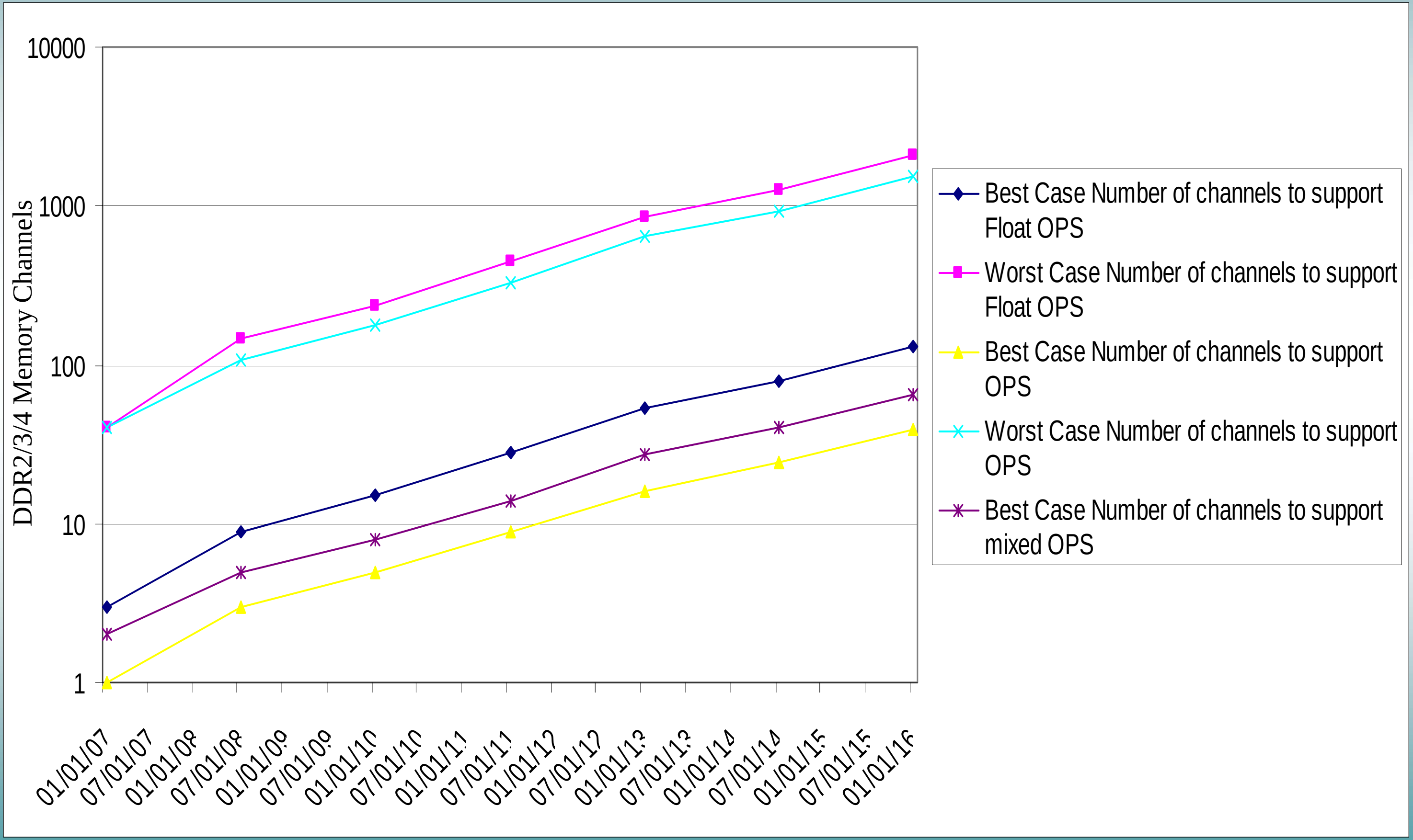
Die photo of Micron's 1-Gbit DDR3, which has a gross-dice estimate of 800 per 300-mm wafer in 78-nm technology.



Die photo of Qimonda's 512-Mbit DDR3, which, at 14.6 x 5.4 mm, is more of a rectangle than the squarish DDR2.

The Bandwidth Crisis:

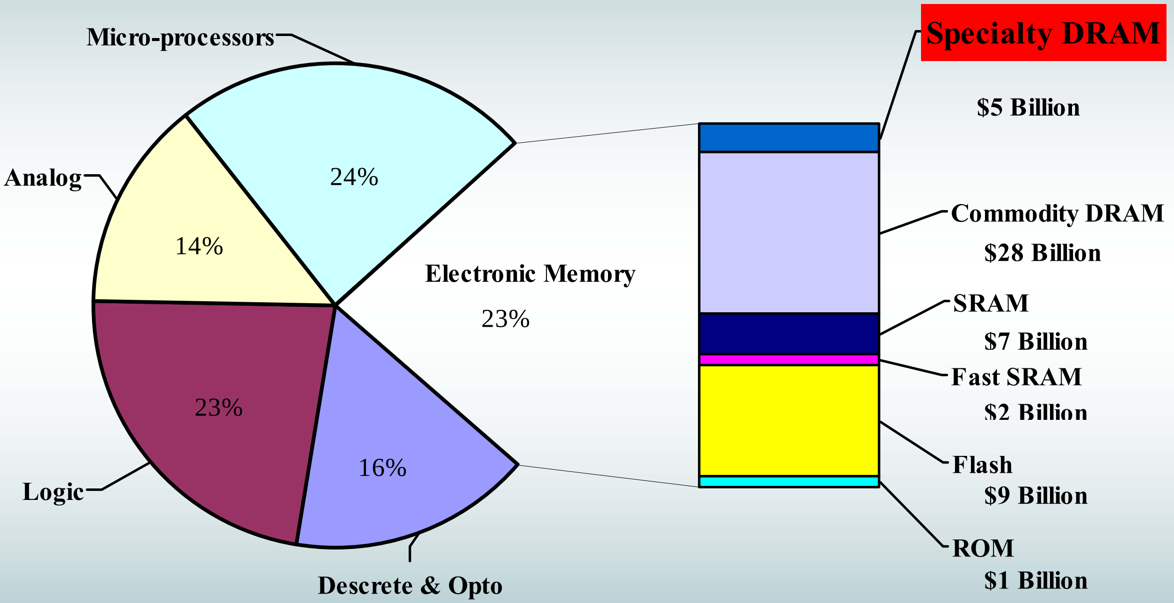
You know you have a problem when there is a log scale….



The Detail

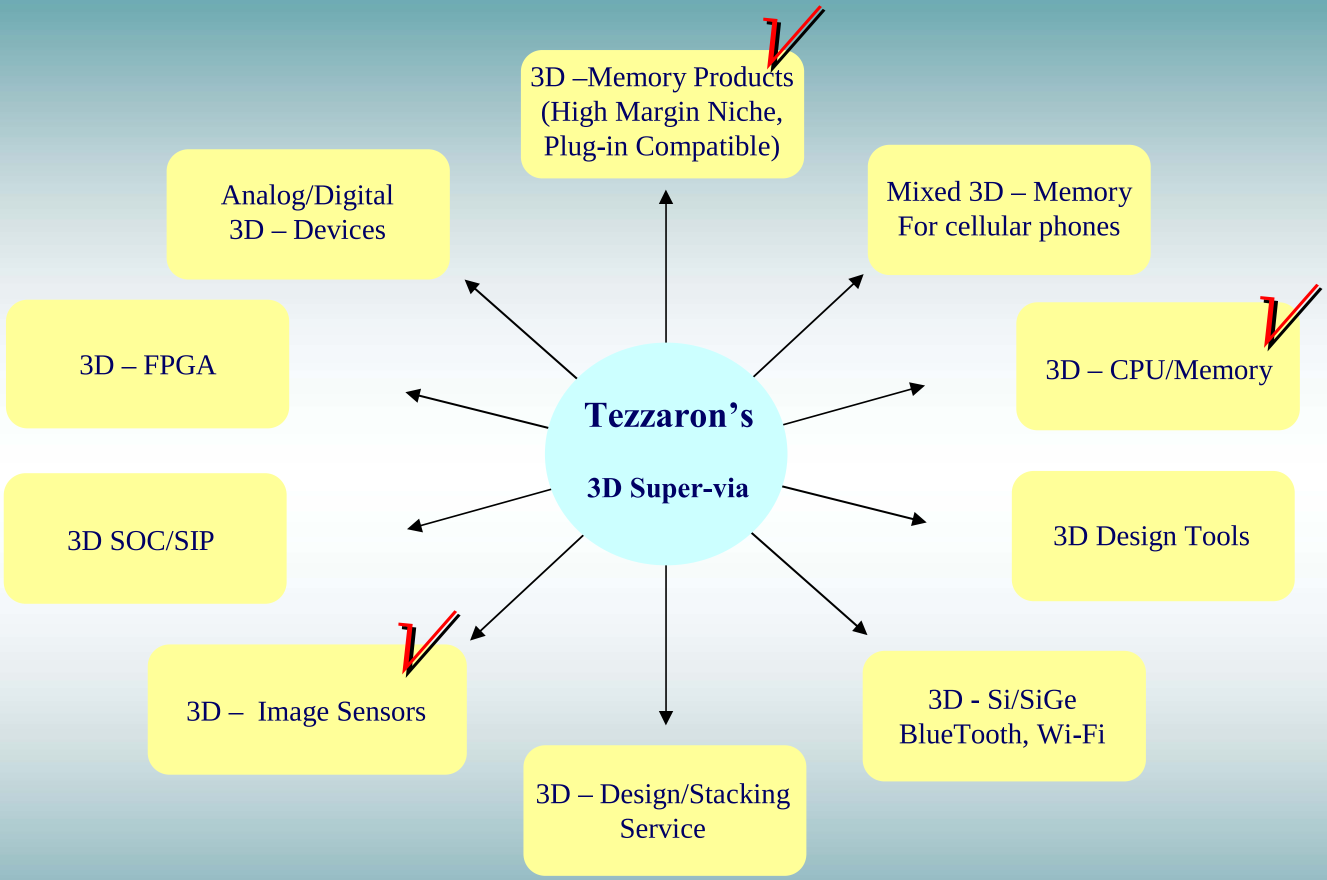
	1/1/2007	7/1/2008	1/1/2010	7/1/2011	1/1/2013	7/1/2014	1/1/2016
Number of cores	4	8	16	32	64	128	256
Clock (GHz)	2	3	3.3	3.63	3.993	4.3923	4.83153
FLOP/core	2	4	4	4	4	4	4
0.5 byte/FLOP (GB/s)	8	48	105.6	232.32	511.104	1124.429	2473.743
8 byte/FLOP (GB/s)	128	768	1689.6	3717.12	8177.664	17990.86	39579.89
OPS/core	4	6	6	6	6	6	6
0.1 byte/OP (GB/s)	3.2	14.4	31.68	69.696	153.3312	337.3286	742.123
0.25 byte/OP (GB/s)	8	36	79.2	174.24	383.328	843.3216	1855.308
4 byte/OP (GB/s)	128	576	1267.2	2787.84	6133.248	13493.15	29684.92
Peak Memory Xfer rate per Channel (GB/s)	6.4	10.7	14.4	16.8	19.2	28.8	38.4
Sustained Memory Xfer rate per Channel (GB/s)	3.2	5.35	7.2	8.4	9.6	14.4	19.2
Best Case Number of channels to support Float OPS	3	9	15	28	54	79	129
Power Required for I/O 40mW/pin (in Watts)	9.6	28.8	48	89.6	172.8	252.8	412.8
Worst Case Number of channels to support Float OPS	40	144	235	443	852	1250	2062
Power Required for I/O 40mW/pin (in Watts)	128	460.8	752	1417.6	2726.4	4000	6598.4
Best Case Number of channels to support OPS	1	3	5	9	16	24	39
Power Required for I/O 40mW/pin (in Watts)	3.2	9.6	16	28.8	51.2	76.8	124.8
Worst Case Number of channels to support OPS	40	108	176	332	639	938	1547
Power Required for I/O 40mW/pin (in Watts)	128	345.6	563.2	1062.4	2044.8	3001.6	4950.4
Best Case Number of channels to support mixed OPS	2	5	8	14	27	40	65
Power Required for I/O 40mW/pin (in Watts)	6.4	16	25.6	44.8	86.4	128	208

Market Size



SIA SEMICONDUCTOR INDUSTRY ASSOCIATION

Addressable 3D Market



Tezzaron's

3D Super-via

3D –Memory Products ✓
(High Margin Niche,
Plug-in Compatible)

Mixed 3D – Memory
For cellular phones

3D – CPU/Memory ✓

3D Design Tools

3D - Si/SiGe
BlueTooth, Wi-Fi

3D – Design/Stacking
Service

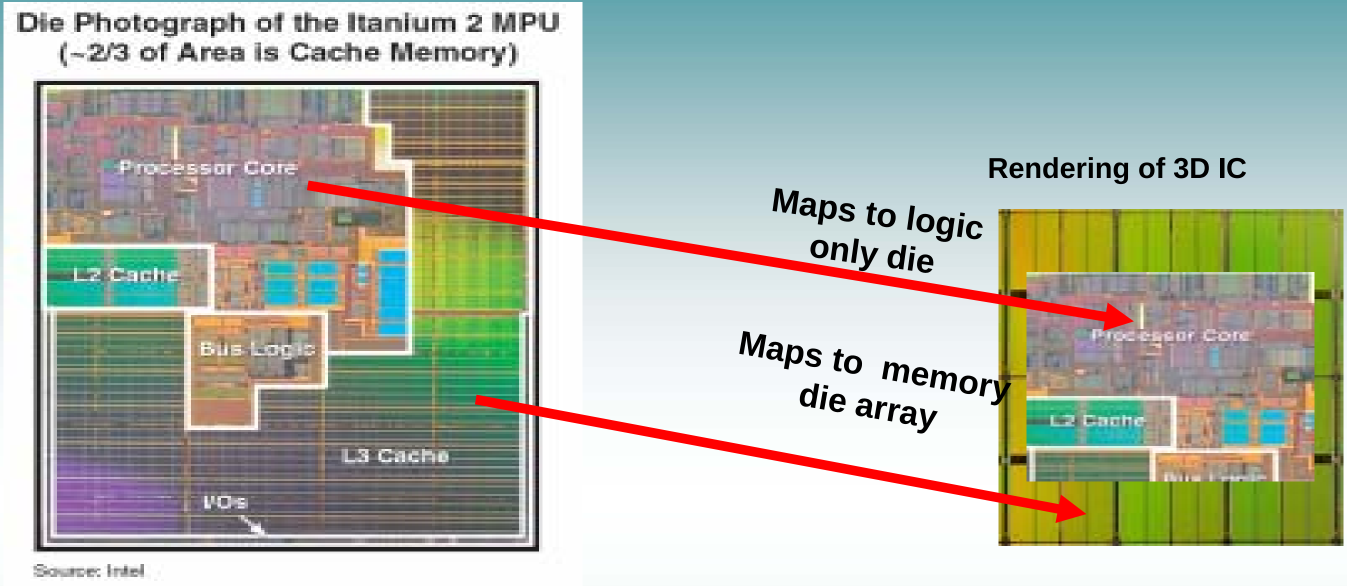
3D – Image Sensors ✓

3D SOC/SIP

3D – FPGA

Analog/Digital
3D – Devices

3D Heterogeneous Integration



Die Photograph of the Itanium 2 MPU (~2/3 of Area is Cache Memory)

Source: Intel

Rendering of 3D IC

Maps to logic only die

Maps to memory die array

BEFORE Intel Photo used as proxy

Only Memory Directly Compatible with Logic (virtually no choice!)

Single Die~ 430 mm2 2D IC “All or Nothing”

Wafer Cost ~ $6,000

Low yield ~ 15%, ~ 10 parts per wafer

memory costs ~ $44/MB

AFTER: 3D IC

14x increase in memory density

4X Logic Cost Reduction

29x → 100x memory cost reduction (choice!)

128MB not 9MB

memory costs ~ $1.50/MB → $0.44/MB