Challenges and Opportunities for Exploiting 3D Technology in System Designs

Mike Ignatowski
IBM Corp
ignatow@us.ibm.com

Photo from April 12 IBM Announcement Of Through Silicon Vias (TSV)

October 11, 2007
This talk will be about systems issues related to exploiting 3D technology

Pat Gelsinger (Intel), Feb 2007

“What you saw as the system architecture yesterday, tomorrow is the on-die architecture.”
Overview

I will try to demonstrate that...

- **3D Integration is an exercise in new system design,**
  - Not just a technology boost

- **Adding 3D technology to a existing system design that was optimized for 2D results in only very modest benefits.**
  - Must rethink processor, memory and I/O designs

- **Power efficiency is a critical issue.**
  - You can integrate more function than you can cost effectively cool.

- **Test and yield issues are a big concern**
  - We must explore yield-tolerant design as a way to mitigate these problems

The potential is huge, but problem must be approached from a system-level perspective to get the full benefit from 3D
Stay honest....
Comparing 3D With Conventional 2D Systems

- Quantification of the near-term benefits of 3DI requires a fair comparison with alternative 2D implementations

**2D Options**

- Chip 1
- Chip 2
- Ceramic / Organic Package

- or -

- Big Chip (1+2)

**3D Options**

- Chip 1’
- Chip 2’
- Silicon Carrier
- Ceramic / Organic Package

- or -

- Chip 1’
- Chip 2’
Typical Evolution of New Technologies

Message: Historically NO new technology produces a 2x-4x improvement when first introduced.
Background

- Traditional CMOS scaling trends
- Some system trends
- The family of 3D technologies
Classical Scaling and Reality

**Interconnect Delays**

- Local (Scaled)
- Global with Repeaters

**Photolithography: Wavelength Longevity**


**Fab cost ($M)**

Source: [www.icknowledge.com](http://www.icknowledge.com)
Impact on Processor Frequency

- High-performance clock speed is saturating
  - Unacceptable Power & variability growth

Source: 2007 IBM GTO
Multi-core Designs Can Help…

- Performance growth has become increasingly dependent on adding cores

**SPECint_2000_rate**

- Single thread performance growth slows significantly
- System throughput continues to grow

Source: 2007 IBM GTO

Note: 55% growth rate refers to “socket” performance
Another Industry Example

Niagara2 Chip Overview
Motivations for exploiting 3D

- Common Quote:

  “2D scaling will slow considerably. When this happens we need to exploit 3D to stay on the performance growth curve”

- This quote turns out to be misleading.

- Even if 2D lithography scaling continues, we will need 3D integration to maintain a 55% CGR (compound growth rate) for performance
  1. Processor frequency is saturating
  2. Lithography scaling of 2x density per 2 years  ->  41% CGR
What is 3D Technology?

- A family of technologies, enabling stacking of active Si layers and vertical connections between them using through silicon vias.
Applications for 3D Technology

- A continuum of technologies and applications exist.

**Via Density (pins/cm²)**

- $10^3$
- $10^4$
- $10^5$
- $10^6$
- $10^7$

**Via Size (μm)**

- 200
- 50
- 10
- 1.0
- 0.20

**Applications**

- Wireless Communications
- Auxiliary Functions
- Hierarchical Cache
- 3D Multicore μP
- 3D Silicon Carrier
- Aggressive 3D Integration
3D Integration: Family of Technologies

Silicon Carriers / Silicon Interposers

Through-Silicon Vias ("TSV")

3DI Assemblies

Layer Bonding and Contacting

Layer Align

High Aspect Ratio Oxide Vias

Chip-Stacking

Flip-Chip Packaging
15 Years of IBM 3D Research

Many programs existed for extended periods before publication

1990
Stacked Memory

1995
Transfer & Join; T1

2000
3D Chains

2005
Annular Via

2010
Cu-Cu Bond

Patent activity 60’s-80’s

1995
Laser Release

1995
TSV Etch

2000
3D Transfer

2000
System on Package Development

2000
Si Carrier w/ OE Cavity

2005
TSV Press Release

Many programs existed for extended periods before publication.
Why should we consider Si Carriers to be “3D technology”?

To build this you must master the following:

1. Technology of making through-silicon vias
2. Manufacturing and testing multi-chip assemblies
3. Changes to the design approach
   - Chip-to-chip wiring can be much denser and lower power than with other forms of packaging.
Challenges to Exploiting 3D Integration
Challenges:
In rough order of priority to design teams

- Manufacturing
- Test
- Yield / Cost
- Reliability
- **Quantify Benefits**
  - Improved Power / Performance
  - Or improved Cost / Performance
  - Or reduced packaging costs
  - Or improved time to market
- **Design Tools and Technology Ground rules**
- Thermal Issues
- Electrostatic Discharge Protection

Reduce risk as much as possible
Testing and Yield Implications

Two alternatives:

1. Test each chip before stacking and again after packaging
   - Test cost doubles!
2. Package the entire stack and then test
   - Does not double test cost, but
   - Yield problems

- Must solve these problems
  - Yield-tolerant chip and system design
Design for Yield and Resiliency

Characteristics of chips with future technologies

- More time dependent degradations and field failures
- Burn-in is becoming more difficult
- Implication - One time factory test will be insufficient

Solution

- Dynamically test and reconfigure around faults in the field
- Spare components will become more standard design features
- 3D enables integrating more components than you can power-on at any one time, so it fits in well with spare component schemes.
Redundancy for all critical components may not be as silly as it once seemed.
Opportunities Enabled by 3D Integration
Some Opportunities

- **DRAM, FLASH**
  - Expected production in 2008-9

- **3D FPGA**

- **ASICS**

- **Client / Graphics**

- **Servers**

Nvidia GeForce 88000 Graphics chip – 128 processors
Potential Systems Benefits of 3D

- Increased device density per unit volume
  - Reduce cost of next level of packaging

- Much higher levels of chip-to-chip I/O connections
  - Higher density, lower latency, and at lower power levels
    
    Chip-to-chip Power < 1mw/Gbps for silicon carrier options, compared to 3-6mw/Gbps for aggressive chip I/O between sockets / boards.

- Ability to tightly integrate dissimilar technologies
  - Different CMOS generations or processes,
  - Such as logic and DRAM, or digital and analog
  - Other materials, such as SiGe or optics

- New mechanisms for power and variability management
  - Dedicated layer for voltage regulators, decoupling capacitors
Addressing the Memory Bandwidth Challenge:
Why Larger Caches are Needed
Multithreaded CPUs: Bandwidth Issues

- Increasing threads per socket requires:
  - increased cache capacity
  - Increased memory bandwidth

- Lack of memory bandwidth growth drives the need for even more cache capacity
  - Example: 2x threads with NO increase in memory bandwidth requires 8x more cache capacity

![DRAM bandwidth chart](image)

55% per year Growth Trend (based on chip performance growth)
Add Multithreading

4 core chip

Add Virtualization

Modern system design
Can be very stressful
On the cache hierarchy

“Looks Like” 4 separate 16 core systems
Options for changing Processor / Cache integration schemes

- Adding 3D elements to an existing design optimized for 2D gives only modest benefits.
- Restructuring the compute node to take advantage of 3D bandwidth and capacity results in more significant performance improvements.
Options for Changing Processors Designs to Exploit 3D

A “Spherical Processor”
- Build processor on several layers to minimize wire lengths and thus minimize cycle time
- Hard to get more than a 10% improvement
- Very hard to test before assembly
- Very hard to debug the hardware

Alternative approach - Intel study by Bryan Black et al.
- Remap a Pentium 4 into two stacked chips at a macro level.
- Don’t try to improve the cycle time, just take out a pipeline stage here and there.

Results
- 15% performance improvement, and a 15% power reduction
  But, using voltage/frequency scaling to get back to equal performance gives:
- Same performance as 2D design, with 64% less power (!)
What about exploiting very high bandwidth to the processors?

- Current processors called “64b” or “32b” for a reason
- Designs and cannot make use of larger blocks of data
- Operating on small data blocks (1-8 bytes) is built right into the instruction set.
- Exploitation of very high bandwidth would require defining a new highly parallel processor (or accelerator) with a new instruction set.
3D Systems: Many Possible Options Exist

- What new systems does 3D enable?
We can design (in PowerPoint) a chip stack with a tremendous amount of computational power.

Are there any other minor problems that might also need attention?

Just a few, if we continue with business as usual, then…

1. **We cannot get enough well regulated power in**
   - And stay within thermal constraints

2. **We cannot get enough data in**
   - And stay within the packaging and power constraints

3. **We cannot provide enough memory bandwidth and capacity**
   - And stay within cost and packaging constraints

4. **We cannot design and validate the chips**
   - And stay within the schedule and resource constraints
Summary of Systems Benefits

- **3D integration can be thought of as:**
  - Offering the performance improvement equivalent to that obtained by one generation of silicon for high-end design
  - At a much lower technology development cost
  - This can be done with a 2-chip 3D stack

- **But is this a one-shot benefit?**
  - Or can it be part of a ongoing long term growth path?
  - The answer is yes, if we invest in the technology to build larger 3D stacks with more chips in subsequent product generation.
Technology / Architecture Transitions

Bipolar → Power problems

CMOS (single chip processor) → Diminishing returns on complexity and frequency

Multi-core / Multi-thread → Cache capacity and Memory BW limits

3D Integrated System → Thermal Limits?