Heterogeneous SoCs Through Advanced Packaging

Subramanian S. Iyer
(s.s.iyer@ucla.edu)

chips.ucla.edu

Our Main Thrust: What are we trying to do?

Develop an “app-like” environment for Hardware that can

• Cut the time to market by 5-10X
• Cut the NRE cost by 10-20X
• Allow extreme heterogeneity including extensions to cyber-physical systems
• Develop a sophisticated manufacturing workforce
The Power of System on a Chip (SoC)

1964 - Transistor
SLT module
6 transistors, 4 resistors

2016-7 – IBM POWER 9 Processor fabbed @ Global Foundries
14nm SOI eDRAM technology, 650mm²
24 cores and 120MB of on-chip eDRAM memory
8 billion transistors, 17 wiring levels !!

Transistor scaling has made this possible

But as SoCs have gotten more complex
• NRE costs have skyrocketed
• Time to market has become huge
• Manufacturing costs have grown
• and yields have plummeted

Courtesy IBM

Computing is becoming heterogeneous

Hierarchical to Data Centric
And heterogeneous
Its no longer feasible to put all this on one die

Courtesy IBM
Although we did try

• Without question the smallest board is a single large die with all the system functions on it

  Amdahl eventually declared the idea would only work with a 99.99% yield, which wouldn’t happen for 100 years.

• Hence the quest for the largest Yieldable Unit – the SoC (reticle size today will yield at 3-15%)

It’s the interconnect density – Stupid!

You can’t scale yourself out of the power gap

You need a different architecture driven by and ability to interconnect more efficicently
What would Yogi Say?

**So, if you can't scale the chip**

**you should try to scale something else!**

---

**Package/Board Features have scaled modestly**

- Silicon features **1000X**
- Board channel
  - We packed more function on die – SoC era
  - Serialization – deserialization techniques allowed high data rates over fewer channels
- Packaging features ~ 4X
  - Boards are large
  - “Serdes” area (>25%) and power (>25%) getting out of hand
Do we need Packages? Short Answer No!*  

Epoxy/glass Board CTE = 16-20x10^-6 /°C

Packages are supposed to:
• Protect the chip – mechanically and thermally – so why do we have CPI issues and heat sinks
• Connect the chip electrically to other chips – really? So why do we have to fan-out?
• Allows mission mode testing of the chip - Ok – I think this one is real!

*I do realize that this is a dangerous thing to say at a packaging conference

Evolution Of System Integration

Interposers/Boards
Organic -> Si-> glass

Massively Integrated Silicon-like Board

Prehistoric
Integration

Stacked Die

Stacked memory

Now

Future

Wafer stacking
Mega SoIFs* by re-integration on an Interconnect Fabric

* System on Interconnect Fabric

CHIPS Interconnect: Fine pitch and small spacings

- Very wide parallel interconnect
  - Routing pitch similar to CMOS top-level routing
  - Bump/pad pitch as fine as 2μm possible

- SERDES can be avoided except for very long links
  - Even then, complexity expected to be lower than traditional SERDES

- Primary benefit: <0.2pJ/b possible even at multi TB/s aggregate data rates
Dielet size and interconnect pitch - Case Study - BlueGene® Q

- Most dielets are small (< 3mm x 3mm)
- Large dielets are dominated by interconnect congestion – can be mitigated by extra wiring levels
- Interconnect pitch of 5 µm would be sufficient

Key constraints: Interconnect pitch and dielet size

- Mechanical constraints
- Die handling constraint
- CMOS wire – like
- Optimal pitch: 2-10 µm
- SoC – like
- Packaging like
- IP re-use
- Die yielding constraint
- Electrical/logical constraints
- I/O complexity
- I/O power
- Testing complexity
- Contacted Gate pitch ~ 50 nm
Mega SoIFs* by re-integration on an Interconnect Fabric

The "right" interconnect fabric
- Mechanically robust (flat, stiff, tough...)
- Capable of fine wiring, fine pitch interconnects
- Thermally conductive
- Can have active and passive built-in components

Silicon Fits the Bill in many cases

Challenges:
- Warpage
- Topography
- Assembly / Thru’put
- Thermal

These issues are largely mitigated by going to a silicon IF with small Dielets

Silicon Fit the Bill in many cases

Full contact – TCB
Proximate
- Inductive
- Capacitive

Preliminary Results

Fig. 10 Cross Section Optical (top) and SEM (middle) images of the bond: The bond is between blank die with 5um Cu layer (coated by 150um Au layer for Cu passivation) and 10um pillar pitch SIF at the location showed at the inset of the images. The bond shows good quality at Cu-Au-Cu bond interface from 12000X SEM image (bottom).

Fig. 11 Top view of the blank Si dies integrated on the SIF covered by Cu layer, showing 100um die-to-die distance.

Fig. 12 A 5mm-by-5mm die integrated on the test SIF (Fig. 9) which has 4um diameter Cu pillars and 10um pitch.
Multiple die on Si Interconnect Fabric

- Dielet size: 4mmX4 mm
- Fanout size 9X9 to 13X13 mm²
- IF size: 100mm dia
- Force 200N
- Temp 350°C
- Time 1’ per dielet

CHIPS Framework

- Applications & Architecture
  - Heterogeneous Systems
  - Approximate Computing
  - Cognitive Computing
  - Fault Tolerance
  - Supply chain Integrity
  - Security
  - Memory Subsystems
  - Processing in Memory
  - DFT
  - Network on Board

- Design Infrastructure
  - Thermo-Mechanical
  - Electrical
  - Tools
  - Partitioning
  - DFT
  - Active IF Design

- Tier 1 Equipment Partners
  - New Tool Concepts
  - Tool Development
  - Scale-Up

- Materials
  - Fine Pitch Interconnect
  - Substrate Materials
  - Wafer, Stress Flexible Materials
  - Thermal solutions
  - Materials for Additive Mfg. Reliability

- Devices/Components
  - Novel switches
  - New memory
  - MEMS
  - Sensors
  - Passives, antennas
  - Medical devices

- Tier 1 Foundry Partners
  - Si, Compound Semis, MEMS, and OSATs

- Integration
  - A processing facility for Interconnect Fabric (IF) & assembly
    - Silicon processing
    - Glass
    - Flexible substrates
    - Additive manufacturing
    - Thermal compression bonding
    - Wafer thinning
    - Wafer-wafer integration
Bio Compatible FlexTrate®

- High-performance heterogeneous devices/components
- A very large die
- An array of small dies with sub-functional blocks
- 615 Si dies embedded in PDMS

UCLA ENGINEERING
Henry Samueli School of Engineering and Applied Science

Conventional Approach

SoC Chip Data

- Data Fragmentation into dielets
- Foundry source verification of dielets
- Reintegration on Si Interconnect Fabric

Approach
Summary

• We are in the midst of a significant hardware transformation
  • Semiconductor Scaling is saturating
  • Computing models are changing
  • Systems are getting more heterogeneous
  • SoCs design costs and Times-to-Market huge

• The CHIPS approach will drive a much more holistic Moore’s Law

• But “it takes a village”
  • Industry Partners
  • University Partners
  • Government Agencies

Acknowledgements

Faculty Profs. Mark Goorsky, Sudhakar Pamarti, Puneet Gupta

Visiting Scholars: Adeel Bajwa, Tak Fukushima

Students: Siva Jangham, Zhe Wan, Saptadeep Pal, Steven Moran, Goutham Ezhilarasu, Faraz Khan, Xuefeng Gu, Prem Kittur, Arsalan Alam, K.T. Kannan

We thank DARPA (N00014-16-1-2639 ) and members of the UCLA CHIPS consortium for their support of this work.

chips.ucla.edu