Signaling for wafer-scale systems

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

Center for Heterogeneous Integration and Performance Scaling
chips.ucla.edu

Discussion with ODSA group on November 20, 2020
UCLA CHIPS

A UCLA Led partnership to develop Applications, Enablement and Core technologies and the eco-system required for continuing Moore’s Law at the Package and System Integration levels and develop our students & scholars to lead this effort.

Simplify hardware development through novel architectures, integration methods, technologies, and devices.
What we do @UCLA CHIPS

- Large Scale Energy Efficient Systems
- Medical Engineering applications

- Advanced Packaging Technologies
- Novel Compute architectures

- Silicon as a heterogeneous fine pitch packaging Platform, Si IF
- FlexTrate as a flexible Biocompatible Heterogeneous Integration Platform
- The CTT as an in-memory compute device
Silicon and Package scaling

Since 2015, Packaging has taken off!

Why?

Advanced packaging borrowed immensely from Silicon technology

Adapted from: S.S. Iyer, MRS bulletin (2015)
Why is heterogeneity assuming sudden importance?

• Packaging has always been about assembling heterogeneous dies/chips onto a Printed Circuit Board

• The problem with PCBs has to do with Latency and Bandwidth between the chips as well as energy per bit transferred
Neural networks are central to AI
Accuracy requires these networks
to be extremely deep (many hidden layers)
Eg. Residual Net (ResNet) has ~1000+ layers
Also the width of these hidden layers can
also be quite large.

Vector multiplications are a key operation in neural networks
And the vector multiply and accumulate (MAC) function is central
The bit precision of the inputs, weights and outputs can exceed 16 bit,
leading to unprecedented computational complexity.

Even with today’s very powerful processors, processors need
to time multiplex, **constantly** moving inputs, weights and
outputs of each layer between the processor and memories
So the memory bottleneck is quite severe.
This is where packaging comes in! - BW, energy-per-bit Xferred
(and latency) define system performance (and processor speed).
Some observations

• If Moore’s law has enabled miniaturization, why have chips gotten larger?
  – More complex problems
  – More cores @ higher clock speeds
  – More cache memory

• Main memory capacity and access limits performance

• Power density challenges - more "dark" silicon

• I/Os take up more space and power as system size increases >30%

Intel Pentium cpu ~300mm²
-3.1 Million Xtors (1993)
0.8 μm technology

NVidia A100: 54 Billion Xtors - 826 mm² (2020)
In TSMC 7 node

17,000 more transistors
Can this be Done practically?

Some more observations:

• Interposers are getting bigger
• 3D stacks are getting taller
• Interposers are an additional level in the packaging hierarchy

Going to a silicon-like board
With fine pitch interconnect and short die to die spacings will allow us to build massive systems

But many issues need to be addressed
The “Right” Rigid Interconnect Fabric

Requirements:

• Mechanically robust (flat, stiff, tough...)

• Processability: fine pitch wiring, & interconnects

• Thermally conductive

• Can have passive (and active) built-in components

<table>
<thead>
<tr>
<th>Material</th>
<th>Young's Modulus Mpa</th>
<th>Tensile strength Mpa</th>
<th>CTE ppm</th>
<th>Thermal Conductivity W/m-K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organic</td>
<td>0.1 to 20</td>
<td>2000-3000</td>
<td>14-70</td>
<td>0.3 - 1</td>
</tr>
<tr>
<td>Glass</td>
<td>50-90</td>
<td>33-3500</td>
<td>4-9</td>
<td>1-2</td>
</tr>
<tr>
<td>Silicon</td>
<td>130-185</td>
<td>5000-9000</td>
<td>3-5</td>
<td>148</td>
</tr>
<tr>
<td>Steel</td>
<td>190-200</td>
<td>400-500</td>
<td>11-13</td>
<td>16-25</td>
</tr>
<tr>
<td>Copper</td>
<td></td>
<td></td>
<td></td>
<td>400</td>
</tr>
</tbody>
</table>

Organic (e.g. FR-4)

Hybrid approaches (EMIB by Intel)

Economical
Going to a silicon wafer scale is not new - there is a new "twist"

Cerebras (2019) - wafer scale AI processor

Our approach:
Integrate lots of dielets on a silicon substrate at fine pitches

Bumped 100 mm wafer (Ca 1982)
Trilogy Systems

Heterogenous Technology
UCLA CHIPS 2019
Wafer scale Heterogeneous assembly
Important Questions

• What is the optimal pitch at which dies should be interconnected?

• What is the optimal dielet size

• How close should we assemble dies

• What level of heterogeneity should we aim for

Hint: how do we make a SOW look like an ginonormous SOC
The CHIPLET Golden Regime

Mechanical constraints

Optimal pitch
2 to 10 µm

Optimal dielet size
1 to 100 mm²

Electrical/logical constraints

Die yielding constraint

SerDes-like

SoC-like

Packaging-like

Die handling constraint

CMOS wire-like

Interconnect pitch

50 nm

50 nm

Gate pitch

Dielet/chiplet size (# of circuits)

IP reuse

I/O complexity/power

Testing complexity

BGA/LGA

500 µm

500 µm

IP reuse

Testing complexity

ODSA 2020

©S.S. Iyer 2020
What is the optimal I/O pitch?

<table>
<thead>
<tr>
<th>Chip</th>
<th>Area (mm²)</th>
<th>Transistor count (x10⁹)</th>
<th>Technology node (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM POWER9 [26]</td>
<td>695</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>AMD Zen [27]</td>
<td>44</td>
<td>1.4</td>
<td>14</td>
</tr>
<tr>
<td>IBM POWER8 [28]</td>
<td>649</td>
<td>4.2</td>
<td>22</td>
</tr>
<tr>
<td>Intel Xeon Haswell E5 [29]</td>
<td>663</td>
<td>5.56</td>
<td>32</td>
</tr>
<tr>
<td>IBM POWER7 + 80 MB [30]</td>
<td>567</td>
<td>2.1</td>
<td>32</td>
</tr>
<tr>
<td>Intel Itanium Poulson [31]</td>
<td>544</td>
<td>3.1</td>
<td>45</td>
</tr>
<tr>
<td>IBM POWER7 + 32 MB [32]</td>
<td>567</td>
<td>1.9</td>
<td>45</td>
</tr>
<tr>
<td>Intel Xeon 7400 [33]</td>
<td>503</td>
<td>1.9</td>
<td>45</td>
</tr>
</tbody>
</table>
Practical limits in heterogeneous integration

- **Fine pitch?**
  - like "fat wires" on a Silicon wafer - 2-10 \( \mu \text{m} \) - this is the bump pitch (BGA pitch is >500\( \mu \text{m} \))
  - Trace pitch < 1 \( \mu \text{m} \) (compared to ~30 \( \mu \text{m} \) on PCB)

- **Precision alignment?**
  - similar to fat wire alignment <0.2 \( \mu \text{m} \) (bumps alignment accuracy is several \( \mu \text{m} \))

- **Close Spacing**
  - As close as possible <20 \( \mu \text{m} \) (dies on a PCB are spaced at least a few 10’s of mm away)

- **Typical block sizes on an SoC are typically a few ~100 \( \mu \text{m} \) on a side**
  - So dielets should be small (1 to 100 \( \text{mm}^2 \) in area)

- **Heterogeneity:**
  - multiple nodes - use the node that is optimal from a performance, area and cost perspective
  - multiple technologies - logic, DRAM, sensors etc.
  - multiple materials Si, III-Vs........
A versatile Fine pitch wafer-scale assembly (Si IF)

Direct Cu-Cu Thermal Compression Bonding using formic acid vapor

X-Ray Tomograph of 10µm Cu-Cu pitch die to wafer connects

Process parameters
- Pressure
- Temperature
- Surface prep

2x2 array of TSMC 16FF dies on the Si-IF.

55 µm inter-die spacing

Legacy dies & passives on Si-IF

Direct Cu-Cu Thermal Compression Bonding using formic acid vapor

Developed termination protocols with most major foundries Cu for Si, Au for III

Wafer scale assembly at fine pitch

Direct Cu-Cu Thermal Compression Bonding using formic acid vapor

Developed termination protocols with most major foundries Cu for Si, Au for III

Legacy dies & passives on Si-IF

Direct Cu-Cu Thermal Compression Bonding using formic acid vapor

Developed termination protocols with most major foundries Cu for Si, Au for III

Legacy dies & passives on Si-IF
Established CHIPS metrics using SuperCHIPS macros

- Continuity check
- Latency characterization
  - Reference & Si-IF ring oscillator: 3-4 GHz
  - On-chip frequency divider ($2^{12}$) & cycle counter
- High-speed data transfer & Bit error rate (BER)
  - Programmable ring oscillator clock: 0.5-3 GHz
  - Pseudo Random Number Generator (PRNG)
  - On-chip comparator and error counter
• Successfully passed continuity tests of both passive and active daisy chains

• Measured latency verified with on-chip counter
  – Latency comparable to on-chip buffer delays
  – Overall latency is <30 ps

• Demonstrated data transfer up to 3 Gbps
  – Bandwidth: 1200 Gbps/mm for 2-layer Si-IF
  – No errors were observed even after 43 hrs
  – BER: <10^{-14} with 99% confidence (Estimate: <10^{-25})

• Measured energy/bit: 0.028 pJ/b

• No electrostatic discharge protection (ESD) used
  – For ESD protection of 50 fF: Latency & Energy increase by

---

<table>
<thead>
<tr>
<th>Oscillator</th>
<th>Measured frequency [kHz]</th>
<th>Actual frequency [GHz]</th>
<th>Latency of Si-IF links [ps]</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSMC 16FF Die</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>On-chip reference</td>
<td>921.1</td>
<td>3.77</td>
<td>NA</td>
</tr>
<tr>
<td>200 μm Si-IF links</td>
<td>836.8</td>
<td>3.43</td>
<td>6.67</td>
</tr>
<tr>
<td>500 μm Si-IF links</td>
<td>762.3</td>
<td>3.12</td>
<td>13.80</td>
</tr>
<tr>
<td>GF 22FDX Die</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>On-chip reference</td>
<td>1033.9</td>
<td>4.23</td>
<td>NA</td>
</tr>
<tr>
<td>200 μm Si-IF links</td>
<td>877.6</td>
<td>3.59</td>
<td>10.51</td>
</tr>
<tr>
<td>500 μm Si-IF links</td>
<td>760.3</td>
<td>3.11</td>
<td>21.26</td>
</tr>
</tbody>
</table>

---

Measured waveforms for TSMC 16FF die assembly
SuperChips - a versatile communication protocol

Micrograph of the fabricated SuperChips interface

SuperChips macros
Width: 1.5μm
Pitch: 4.9μm

Cu pillars (Ø = 4 μm)

9.8 μm
350 μm

8111 I/O interdie Connections
22291 power Connections

Schematic of the SuperChips I/O

Async_Sel
Data_in
Clock_in

D Q

Q

Si-IF link

D Q

Q

Data_out

Longer Range connections can be done daisy chaining through intervening dies using porosity rules and multiple buffer stages - for a few die over or using pico-SerDes for longer (~ cms) lengths.

Using “utility dies” which may also provide redundant routing options to manage assembly defects

Technology/Interface protocol

<table>
<thead>
<tr>
<th></th>
<th>Si-IF/ SuperChips</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Async</td>
</tr>
<tr>
<td>Interconnect pitch</td>
<td>10 μm</td>
</tr>
<tr>
<td>Overall Latency (ps)</td>
<td>30</td>
</tr>
<tr>
<td>Data-rate/link (Gbps)</td>
<td>10</td>
</tr>
<tr>
<td>Energy/bit (pJ/b)</td>
<td>&lt;0.03</td>
</tr>
<tr>
<td>Maximum Bandwidth/mm (Gbps/mm)</td>
<td>8000^a</td>
</tr>
</tbody>
</table>

^a
^b
## Technology Comparison using s-FOM$_k$

<table>
<thead>
<tr>
<th>Tech/Interface protocol</th>
<th>Si-IF/SuperCHIPS</th>
<th>Interposer r/AIB</th>
<th>PCB/SerDes</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Async</td>
<td>Sync</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reach</td>
<td>Neighbor</td>
<td>Neighbor</td>
<td>Neighbor</td>
<td>Long Reach</td>
</tr>
<tr>
<td>Overall Latency (ps)</td>
<td>30</td>
<td>500</td>
<td>1500[1]</td>
<td>~2000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>~6000</td>
<td>3-65X</td>
</tr>
<tr>
<td>Energy/bit (pJ/b)</td>
<td>&lt;0.03</td>
<td>&lt;0.15</td>
<td>0.8-0.85[3,4]</td>
<td>1.17[7]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6.9[13]</td>
<td>5-40X</td>
</tr>
<tr>
<td>Bandwidth/mm (Gbps/mm)</td>
<td>8000$^a$</td>
<td>2560$^{a,b}$</td>
<td>707.7$^b$</td>
<td>354</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>149-298$^c$</td>
<td>4-23X</td>
</tr>
</tbody>
</table>

---

$$s - FoM_k = \frac{\text{Bandwidth} \times \text{Energy/bit}}{\text{mm}}$$

---

* a 4 wiring levels, $^b$ Assuming 20% overhead, $^c$ Estimated from data in [10-13]

---

Technology Comparison s-FoM_{ucla} - shows the benefit of technology

- Does not account for area used by I/Os
  - SerDes occupy significant chip area
  - Especially when we have deep I/Os that go several layers in
  - This can be >30% of die area!
  - Note: this is influenced by Technology node

- Does not account for latency
  - ToF is not always the main contributor
  - Serialization, Deserialization, equalization, clock recovery etc. are the major contributors
  - Note: this is influenced by Circuit design

- No credit for load that is driven
  - This is influenced by Packaging Technology

\[
s - FoM_u = \frac{\text{Bandwidth}_{\text{shoreline}} \times \text{IOcols}}{\text{Energy}_{\text{bit}}} \times \frac{\text{Length}_{\text{link}}}{\text{TransceiverArea}_{\text{Link}}} \times \text{Latency}
\]

This is the die area “wasted” by I/Os and can’t be used for compute

Overhead time to serialize/deserialize data, ECC + ToF

©S.S. Iyer 2020
## CHIPS Project Goals and Milestones

<table>
<thead>
<tr>
<th>Metric</th>
<th>Phase 1</th>
<th>Phase 2</th>
<th>Phase 3</th>
<th>SuperCHIPS on Si-IF (current)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Design level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IP reuse</td>
<td>&gt; 50% public IP blocks</td>
<td>&gt; 50% public IP blocks</td>
<td>&gt; 50% public IP blocks</td>
<td>Feasible</td>
</tr>
<tr>
<td>Modular design</td>
<td>-</td>
<td>-</td>
<td>&gt; 80% reused, &gt; 50% prefabricated IP</td>
<td>Feasible</td>
</tr>
<tr>
<td>Access to IP</td>
<td>&gt; 2 sources of IP</td>
<td>&gt; 2 sources of IP</td>
<td>&gt; 3 sources of IP</td>
<td>2 sources of IP</td>
</tr>
<tr>
<td>Heterogeneous integration</td>
<td>&gt; 2 technologies</td>
<td>&gt; 2 technologies</td>
<td>&gt; 3 technologies</td>
<td>Feasible</td>
</tr>
<tr>
<td>NRE reduction</td>
<td>-</td>
<td>&gt; 50%</td>
<td>&gt; 70%</td>
<td>Feasible</td>
</tr>
<tr>
<td>Turnaround time reduction</td>
<td>-</td>
<td>&gt; 50%</td>
<td>&gt; 70%</td>
<td>Feasible</td>
</tr>
<tr>
<td>Performance benchmarks (performer defined)</td>
<td>-</td>
<td>&gt; 95% benchmark</td>
<td>&gt; 100% benchmark</td>
<td>See s-FeM$_5$</td>
</tr>
<tr>
<td><strong>Digital interfaces</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data-rate (scalable)</td>
<td>10 Gbps</td>
<td>10 Gbps</td>
<td>10 Gbps</td>
<td>10 Gbps</td>
</tr>
<tr>
<td>Energy efficiency</td>
<td>&lt; 1 pJ/bit</td>
<td>&lt; 1 pJ/bit</td>
<td>&lt; 1 pJ/bit</td>
<td>&lt; 0.4 pJ/bit</td>
</tr>
<tr>
<td>Latency</td>
<td>≤ 5 nsec</td>
<td>≤ 5 nsec</td>
<td>≤ 5 nsec</td>
<td>≤ 0.1 nsec</td>
</tr>
<tr>
<td>Bandwidth density</td>
<td>&gt; 1,000 Gbps/mm</td>
<td>&gt; 1,000 Gbps/mm</td>
<td>&gt; 1,000 Gbps/mm</td>
<td>&gt; 1,000 Gbps/mm</td>
</tr>
<tr>
<td><strong>Analog interfaces</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insertion loss (across full bandwidth)</td>
<td>&lt; 1 dB</td>
<td>&lt; 1 dB</td>
<td>&lt; 1 dB</td>
<td>&lt; 0.6 dB at 30 GHz (measured)</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>≥ 50 GHz</td>
<td>≥ 50 GHz</td>
<td>≥ 50 GHz</td>
<td>≥ 50 GHz</td>
</tr>
<tr>
<td>Power handling</td>
<td>≥ 20 dBm</td>
<td>≥ 20 dBm</td>
<td>≥ 20 dBm</td>
<td>≥ 20 dBm (EM limited)</td>
</tr>
</tbody>
</table>
So What are the issues?

• Developing the assembly technology: fine pitch, close spacing, tight alignment etc...
• Establishing a communication protocol for both near and far dielets
• Communicating with the outside world
• Delivering power - huge amounts of power!
• Extracting heat - huge amounts of heat!
• Making such system reliable
• Ensuring the costs are economical
Communicating with the outside world

- Flexible high speed wired connectors (FlexTrate™)
- RF links using embedded fused quartz or PDMS and III-V drivers
- Photonic Interconnects
Summary

• Packaging has scaled significantly in the last few years
  – Driven by need, more investment, More silicon-like processing
  – Silicon as a base packaging material has significant potential

• The challenges are
  – Assembly - especially at high throughput
  – Connections to the outside world
  – Power delivery and heat extraction
  – Reliability and yield
  – Supply chain for bare dies

• We can extend this concept to flexible hybrid electronics (did not talk about it much today)
Selected Bibliography (more [here](#))

- S. Jangam and S. S. Iyer, "A Signaling Figure of Merit (s-FoM) for Advanced Packaging," in IEEE Transactions on Components, Packaging and Manufacturing Technology, doi: 10.1109/TCPMT.2020.3022760
UCLA CHIPS  Thanks you for your Support !
In a recent zoom Group meeting, our students shared updates and insights.