To Chiplet or Not To Chiplet: Heterogeneous Integration and Chiplets

Mudasir Ahmad

IEEE EPS Distinguished Lecturer Talk
March 2022
Acknowledgement

The work presented in this talk is a product of a joint collaboration with ARM and Microchip as part of the OCP/ODSA/Business Work Group

Special Acknowledgement:

Javier DeLaCruz (ARM)

Anu Ramamurthy (Microchip)
Silicon Process Node Challenges

- Development cost of a 5nm device exceeds the cost of the previous two nodes (10nm and 7nm) combined

- With increasing cost of miniaturization, newer silicon nodes need to be used more strategically

- For advanced devices, silicon reticle size is also limiting performance
What is a Chiplet?

Chiplet can be defined as a die specifically designed and optimized for operation within a package in conjunction with other chiplets. This differs from a conventional die which has the drive strength available to enable signaling over a longer electrical distance. A chiplet would not normally be packaged separately and still operate effectively [OCP/ODSA/BWG].
Finding Optimal Chiplet Applications

Printed Circuit Board Assembly

- Packaged components on a PCB
- Need to be miniaturized to:
  - Reduce space
  - Increase performance
  - Reduce cost
  - Scale faster

Chiplets

Monolithic Silicon Blocks

- Blocks in monolithic silicon
- Need to be broken out to:
  - Reuse IP
  - Combine old/new silicon nodes
  - Avoid exceeding reticle limit

Two distinct trends are driving towards Heterogeneous Chiplet Integration
### Chiplets: Benefits

<table>
<thead>
<tr>
<th>Benefit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower Development Cost</td>
<td>Per previous slide, by utilizing older nodes along with newer ones, development cost can be significantly reduced</td>
</tr>
<tr>
<td>Wafer Cost Reduction</td>
<td>Since newer silicon nodes cost more, selective use of advanced nodes can reduce cost overall</td>
</tr>
<tr>
<td>Die Yield Improvement</td>
<td>With newer silicon nodes, smaller dies produce higher yield. E.g. Splitting a 777mm² die into 4 can improve yield by 2.26X*</td>
</tr>
<tr>
<td>Increased Time-to-Market</td>
<td>With IP-reuse and by combining older silicon process nodes, the development time can be reduced, which reduces the time-to-market</td>
</tr>
<tr>
<td>Lower Thermal Cooling Costs</td>
<td>Splitting high heat generating dies can direct focus on cooling the most critical components. Smaller newer node dies could reduce cooling burden</td>
</tr>
<tr>
<td>Lower Re-spin Cost</td>
<td>Re-spinning a new silicon process node can be very expensive. By reusing older nodes and IP, the newer portion that may require re-spins is reduced</td>
</tr>
<tr>
<td>Lower Demand Volatility</td>
<td>The same chiplets can be used across multiple product skews and combined with other chiplets, which smooth out demand volatility for any specific chip</td>
</tr>
</tbody>
</table>

*https://community.arm.com/arm-research/b/articles/posts/three-dimensions-in-3dic-part-1*
## Chiplets: Challenges

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased Scrap Cost (KGD)</td>
<td>If 1 chiplet fails in a 4-chiplet package, does the whole device get scrapped? Cost of test and scrap could increase significantly</td>
</tr>
<tr>
<td>Increased Test Cost</td>
<td>To minimize scrap loss due to KGD, test costs increase and the overall test architecture may need to be updated to “shift left” – increasing cost</td>
</tr>
<tr>
<td>Decreased Package Yield</td>
<td>With multiple chips mounted on a package, packaging complexity increases and this could result in reduced packaging yield</td>
</tr>
<tr>
<td>Business Model</td>
<td>With different suppliers providing different components in a chiplet package, who is responsible for what? Who takes what liability for failures?</td>
</tr>
<tr>
<td>Horizontal/Vertical Scaling Limits</td>
<td>Placing 2 – 4 chiplets is one thing, but with &gt; 20 chiplets, there are physical and package scaling limits – impacting timeline, cost and feasibility</td>
</tr>
<tr>
<td>IP Interface Cost, Timelines</td>
<td>Chiplets need IP to communicate. IP development timelines, cost and ownership issues can become critical bottlenecks</td>
</tr>
<tr>
<td>Performance Degradation</td>
<td>Breaking a monolithic die into two for cost reasons has a performance penalty, especially if the monolithic die is smaller than the reticle size</td>
</tr>
</tbody>
</table>
Key Challenge in Developing Chiplets

- Business Model
- When does it make sense to build Chiplets?
- Which combination of Chiplets makes the most economic sense?
- What parameters need to be tuned to ensure that the Chiplets make economic sense?
- How to handle future uncertainty: pricing, IP availability, yield, product/market fit, Opex, Capex?
- Chiplets involves multiple suppliers working together. How to agree on yield targets?

Industry needs a common Economic Model to address these questions.
Economic Model Concept

- Data remains private, model remains open
- Allows multiple stakeholders to operate with the same assumptions and targets
- Captures multiple factors in varying levels of complexity
- Not too detailed, not too coarse: Enough for early decision making
- Customizable as needed
- Allows for first order evaluation of key influencing factors
- Can be scaled and built upon by all users
Model Features

- Phased Model Development:
  - Phase 1: Spreadsheet-based: rough model and a detailed model
  - Phase 2: Python script-based: account for uncertainties, sensitivity analysis etc.

- Allows for user to override inputs – color coded to indicate this

- Ability to integrate different process nodes

- Allows the die to be integrated with passive silicon (like a silicon interposer)

- Can integrate up to 40 chiplets; can also use this for monolithic designs

- Yield for each process node calculated based on the defect density

- Bose-Einstein model or Murphy models included

- Assembly sequence and scrap costs captured in the model
Key Driver Factors

- Yield
- Quality
- Material
- IP
- Assembly Sequence
- KGD
- Test
- Assembly
- Operations
- NRE
- Time
- Mask Set

24+ variables included
## Model Structure

### Model Input
- 40+ Chips/Chiplets
- Active or Passive Silicon
- Silicon Node
- Wafer Pricing, Yield
- Substrate
- Forecast Volumes
- Assembly Sequence
- Yield, KGD
- IP Cost
- ASP
- … 24+ Variables

### Model Output
- Graphical Output
  - Line Chart
  - Pie Chart
- Pairwise Scenario Comparisons
- Total Cost and Unit Cost ($)
- 5 Year Projections
- % Contribution of:
  - BoM, KGD, Test, NRE, IP etc.
- Gross Margin ($, %) for given ASP
Test Case Examples
Test Cases

- **Test Case 1**: Monolithic vs. chiplet options for standard single die Flip Chip BGA
  - 12% extra silicon area
  - 32% less expensive

- **Test Case 2**: Integrating 16 chiplets onto a substrate
  - 12% extra silicon area
  - 20% less expensive

- **Test Case 3**: Large ASIC integrated with 4 HBM vs. ASIC divided into 2 chiplets integrated with 4HBMs
  - 8% extra silicon area
  - 30% less expensive
Scenario 1

- Cost wise, should we split a monolithic die into two smaller ones?
- Several input parameters assumed in model
- No interposer

<table>
<thead>
<tr>
<th>Option 1</th>
<th>Option 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>65 x 65mm</td>
<td>65 x 65mm</td>
</tr>
<tr>
<td>5nm Die 25.8 x 25.8 mm</td>
<td>5nm Die 14.5 x 25.8 mm</td>
</tr>
<tr>
<td>5nm Die 25.8 x 25.8 mm</td>
<td>5nm Die 14.5 x 25.8 mm</td>
</tr>
</tbody>
</table>

12% Extra SoC Area
Scenario 1: Model Results

- Chiplets are **32% less expensive** than monolithic.
- **Cost Drivers:**
  - Material > KGD > NRE > Misc Cost > IP Interface > Operating Cost > Quality
- Chiplets in this case make economic sense.
Scenario 2

- Cost wise, should we split a monolithic die into two smaller ones?
- Several input parameters assumed in model
- No interposer
- Wafer yield goes down over time, volumes go up over time

Option 1

- 16X
- 16nm
- 6.7 x 5mm

Option 2

- Extra 12% SoC Area
- 5nm Die 25.8 x 14.5 mm
- 5nm Die 25.8 x 14.5 mm
Scenario 2: Model Results

- Chiplets are **20% less expensive** than the monolithic version.
- Cost Drivers:
  - Material > NRE > KGD > Misc. Cost > IP Interface
  - Operating Cost > Quality
- Main die split in this case makes economic sense.
Scenario 3

- Cost wise, should we split a monolithic die into two smaller ones?
- Several input parameters assumed in model
Scenario 3: Model Results

- Chiplets-with-HBM is about **30% less expensive** than monolithic-with-HBM.
- Cost Drivers:
  - Material > KGD > IP Interface > NRE > Misc. Cost
  - Main die split in this case makes economic sense.
Effect of Uncertainty
Future is always uncertain

• The previous cases assume that all the factors are known and fixed

• In real life, several factors are uncertain and could vary wildly. For example:
  
  • **Wafer Yield** could change differently over time
  • **Forecast Demand** / shipment volumes can also change significantly over time
  • **Assembly Yield** also varies and is difficult to predict upfront

• These factors need to be accounted for, to determine if the benefits are still worthwhile

• The model needs to capture all these uncertainties to show how robust the choice is

• Sensitivity analysis can also be performed to determine which uncertain factors have the highest impact – they can then be further analyzed and controlled
Scenario 2 with Uncertainty

- Cost wise, should we split a monolithic die into two smaller ones?
- Several input parameters assumed in model
- No interposers
### Scenario 2 with Uncertainty: Inputs

**27 uncertain inputs:**
- Assembly Yield for each option
- Defect Density (DD) for each option, each year
- Forecast Demand (FD) for each option, each year

**27 distributions:**
- For illustrative purposes, assumed Defect Density and Forecast Demand to be normally distributed
- Assumed Assembly Yield for each option to follow a uniform distribution
- Values used are simply for illustrative purposes and could be dialed in for each life scenarios
Scenario 2 with Uncertainty: Results

<table>
<thead>
<tr>
<th>Cost Benefit</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Year 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worst Case (5% Probability)</td>
<td>$10</td>
<td>$55</td>
<td>$36</td>
<td>$31</td>
<td>$19</td>
</tr>
<tr>
<td>Average (50% Probability)</td>
<td>$156</td>
<td>$82</td>
<td>$57</td>
<td>$48</td>
<td>$33</td>
</tr>
<tr>
<td>Best Case (5% Probability)</td>
<td>$325</td>
<td>$110</td>
<td>$79</td>
<td>$64</td>
<td>$47</td>
</tr>
</tbody>
</table>

Observations:

- In all cases, the Chiplet option is lower cost than the Monolithic case.
- The best and worst case savings can be significantly different from the average values.
- The best case savings are highest in the first few years and they gradually decrease over time.
  - This is because the defect density typically improves over time for any silicon process node.
  - With higher shipment volumes, development costs and IP reuse, even a smaller benefit can be quite significant.
Scenario 2 with Uncertainty: Drivers

In the first year, the key drivers are:

- Monolithic Die Defect Density
- Split Die Defect Density
- Assembly Yield of Option 1 [Monolithic]
- Forecast Demand
Scenario 2 with Uncertainty: Drivers

In the fifth year, the key drivers are:

- Split Die Defect Density
- Assembly Yield of Option 2 [Monolithic]
- Assembly Yield of Option 1 [Chiplets]
- Forecast Demand
Future Work

• Phase 1a [Completed]:
  • Simplified “toy” model added for beginners
  • More use cases and model vetting
  • Addition of more variables
  • Some default values per inputs from contributors

• Phase 1b:
  • Detailed Symbolic Math Algorithm details planned be published

• Phase 2:
  • Python script-based model to capture more factors
  • Sensitivity analysis of critical factors: yield, pricing, volumes, etc.
  • Script could be incorporated into EDA tools for iterative design tradeoff analysis
Symbolic Math [to be published]

\[ W_{fa} = \left( W_{fa} - 60 \pi \left( \frac{W_{fa}}{45} - \frac{1}{3} \right) \right) \]  
Gross Die per Wafer

\[ \text{Price per Wafer (t > 1)} \]

\[ \text{Yield}_{\text{Murphy},i,t} = \left[ 1 - e^{-d_{i,t} \frac{\text{Eff}_{i,t} \times \text{Yield}_{\text{Murphy},i,t}}{d_{i,t} \text{Eff}_{i,t} \times 0.01}} \right] + \text{REC}_{i,t} \]  
Yield of each wafer i at time t (Murphy)

\[ \text{FUP}_{\text{Murphy},i,t} = \left( \frac{P_{\text{wafer},i,t}}{\text{INT}_{\text{GDPW1}} \times \text{Yield}_{\text{Murphy},i,t}} \right) + \text{Probe Cost} \]  
Forecast Unit Price per die i at time t, (Murphy)

\[ \text{MatCost}_{\text{Murphy},i,t} = \text{FUP}_{\text{Murphy},i,t} \times FD_{i,t} \]  
Material Cost of die i at time t (Murphy)

\[ \text{SubsUnitCost}_t = \frac{\text{Subs Unit Price}_t}{\text{Subs Yield}_t} \]  
Substrate Unit Cost at time t

\[ \text{SubsCost}_t = \text{SubsUnitCost}_t \times \text{Subs Demand}_t \]  
Substrate Total Cost at time t

\[ \text{Misc Cost}_t = \text{Package Assy Cost}_t \times FD_{i,\text{substrate}} \]  
Miscellaneous Cost at time t

\[ \text{Asy Yield}_i = \text{Asy Per Step Yield}^{\text{Asy}}_{i,t} \times (1 - \text{Asy Yield}_i) \]  
Yield at each assembly step

\[ \text{Asy Scrap}_{\text{Murphy},i,t} = \sum_{i,t} \text{MatCost}_{\text{Murphy},i,t} \times \text{Asy Yield}_i \times (1 - \text{Asy Yield}_i) \]  
Assembly Scrap at time t (Murphy)

\[ \text{Total IP Cost}_t = \text{IP Cost}_t + \left( \frac{\text{IP Cost per ASP & ASPs}}{100} \times FD_{i,t} \right) \]  
IP interface Cost at time t

\[ \text{Yield Quality Cost}_{\text{Murphy},i,t} = \frac{\text{Field Quality FR}_{i,t} \times 1000000}{\sum_{i,t} \text{FD}_{i,t}} \]  
Field Quality Cost of die i at time t (Murphy)

\[ W_{i} = (W_{S_{i}} + W_{S_{i}}) \times \left( \sum_{i} W_{S_{i}} \times (1 - \text{Yield}_{i} \times W_{S_{i}}) \times \sum_{i} W_{S_{i}} \right) \]  
Wafer Sort if Applicable

\[ W_{f} = (F_{T_{i}} + F_{T_{i}}) \times (1 - \text{Yield}_{i} \times W_{S_{i}} \times F_{T_{i}}) \times \sum_{i} F_{T_{i}} \]  
Final Test
Call to Action

• Model will be released periodically and latest version will be listed on the ODSA website: https://www.opencompute.org/wiki/Server/ODSA (Chiplet Cost Model)

• Reach out to us:
  • If you are interested in developing the cost model
  • If you have an existing set of parameters that could be included
  • If you can help us share and evangelize the concept of an open cost model