To Chiplet or Not To Chiplet: Heterogeneous Integration and Chiplets

Mudasir Ahmad

IEEE EPS Distinguished Lecturer Talk
March 2022

Acknowledgement

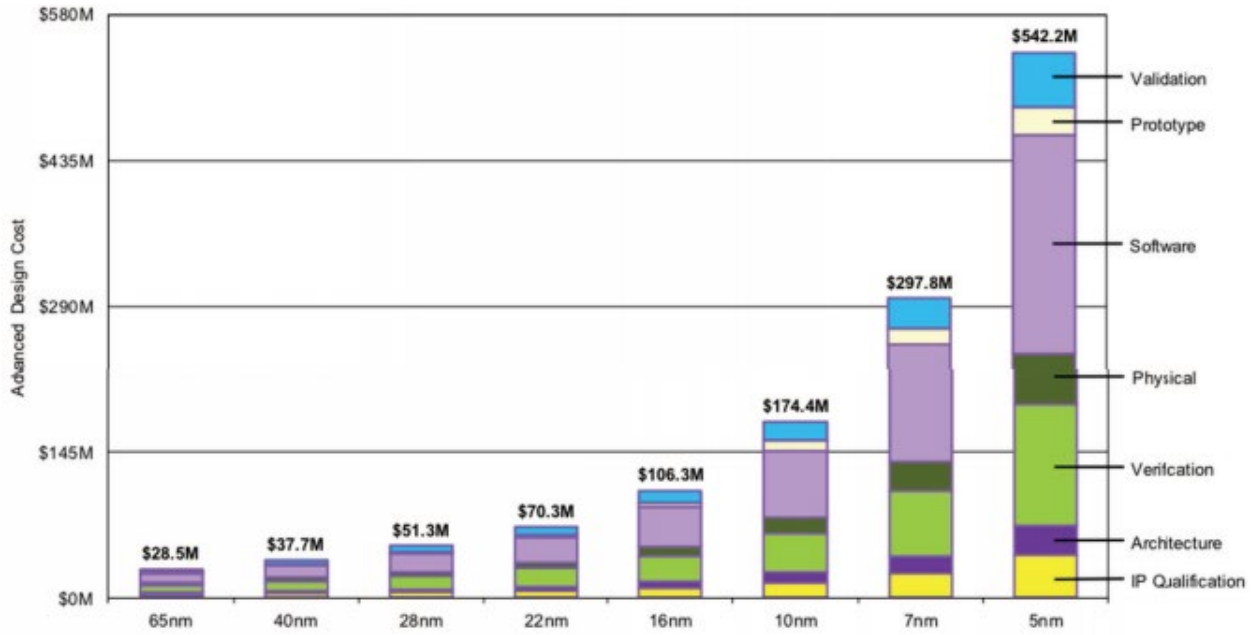
The work presented in this talk is a product of a joint collaboration with ARM and Microchip as part of the OCP/ODSA/Business Work Group

Special Acknowledgement:

Javier DeLaCruz (ARM)

Anu Ramamurthy (Microchip)

Silicon Process Node Challenges



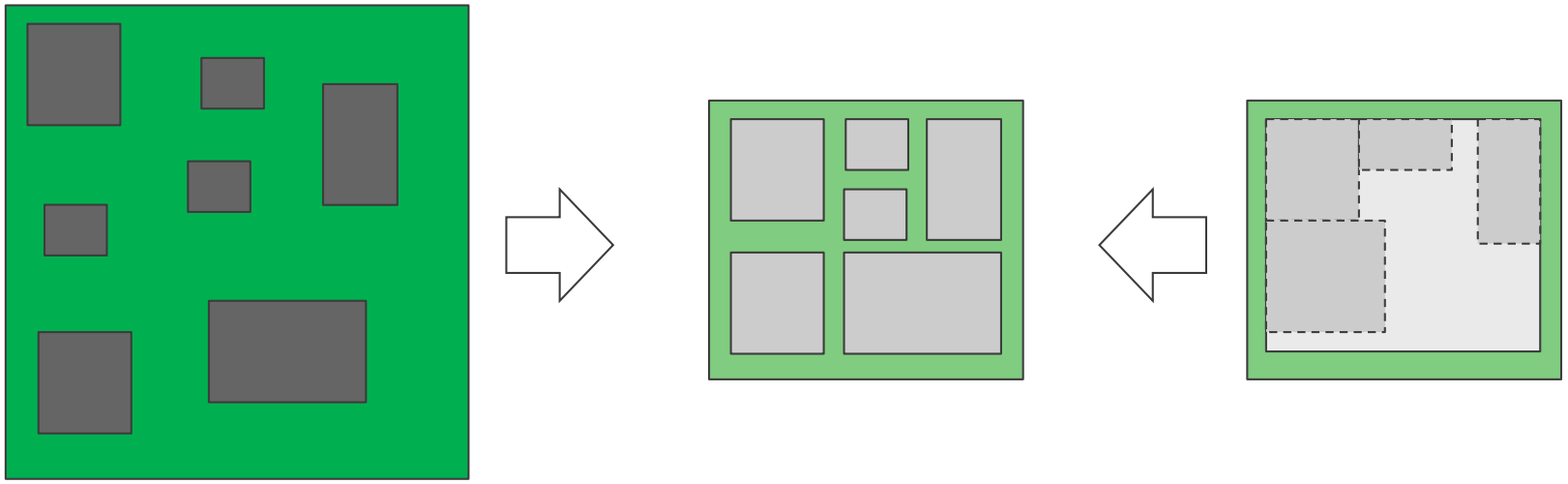
IBS

- 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 | **Chiplets** | **Monolithic Silicon Blocks**

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

- 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

Google

Lower Development Cost	Per previous slide, by utilizing older nodes along with newer ones, development cost can be significantly reduced
Wafer Cost Reduction	Since newer silicon nodes cost more, selective use of advanced nodes can reduce cost overall
Die Yield Improvement	With newer silicon nodes, smaller dies produce higher yield. E.g. Splitting a 777mm^2 die into 4 can improve yield by 2.26X*
Increased Time-to-Market	With IP-reuse and by combining older silicon process nodes, the development time can be reduced, which reduces the time-to-market
Lower Thermal Cooling Costs	Splitting high heat generating dies can direct focus on cooling the most critical components. Smaller newer node dies could reduce cooling burden
Lower Re-spin Cost	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
Lower Demand Volatility	The same chiplets can be used across multiple product skews and combined with other chiplets, which smooth out demand volatility for any specific chip

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

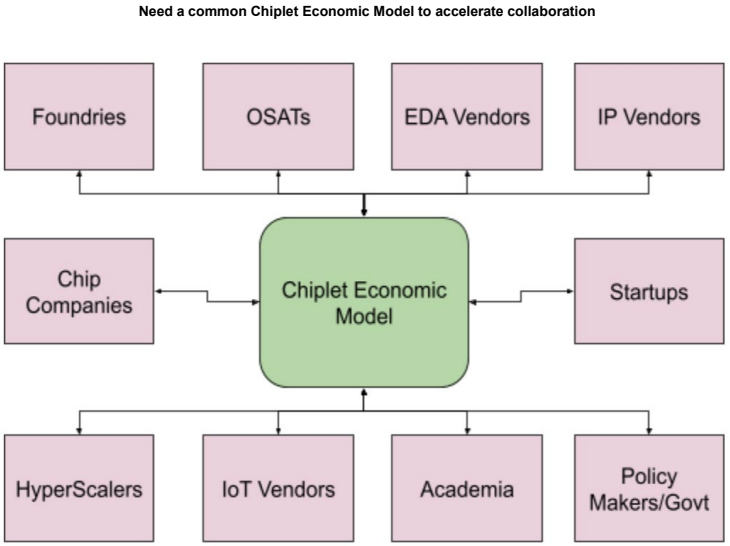
Chiplets: Challenges

Google

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

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

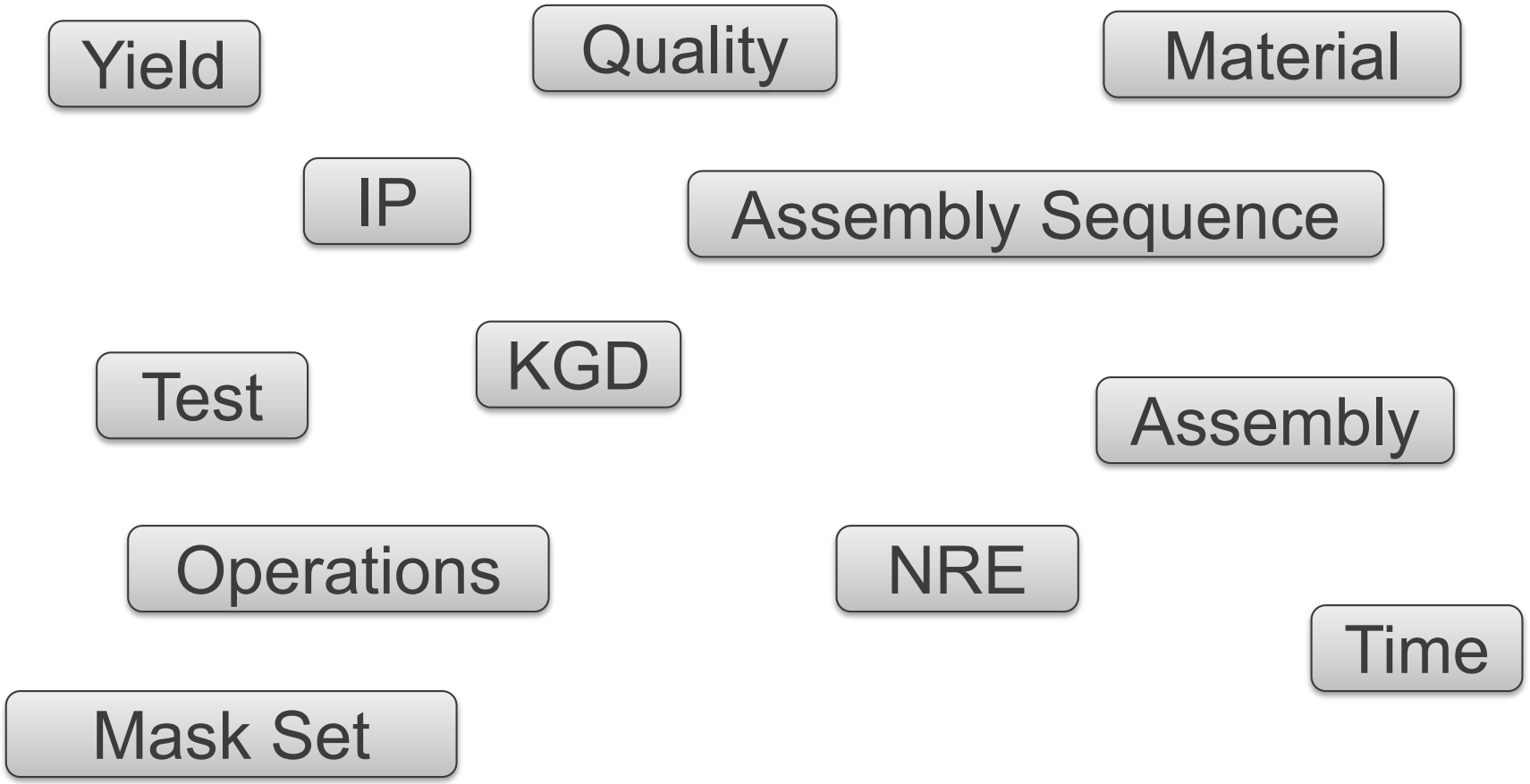
Economic Model Concept

- 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



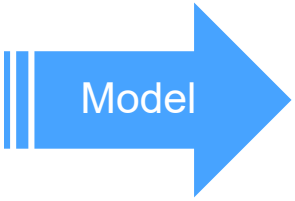
24+ variables included

Model Structure

Google

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



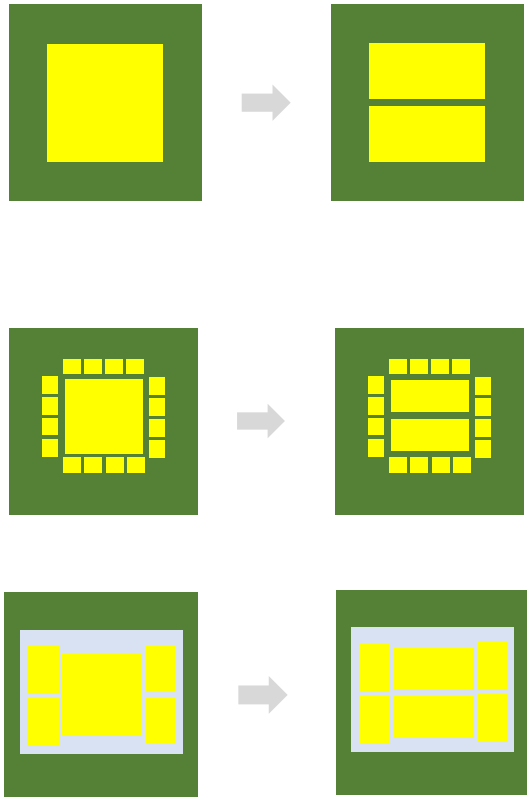
Model Output

- 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

Google

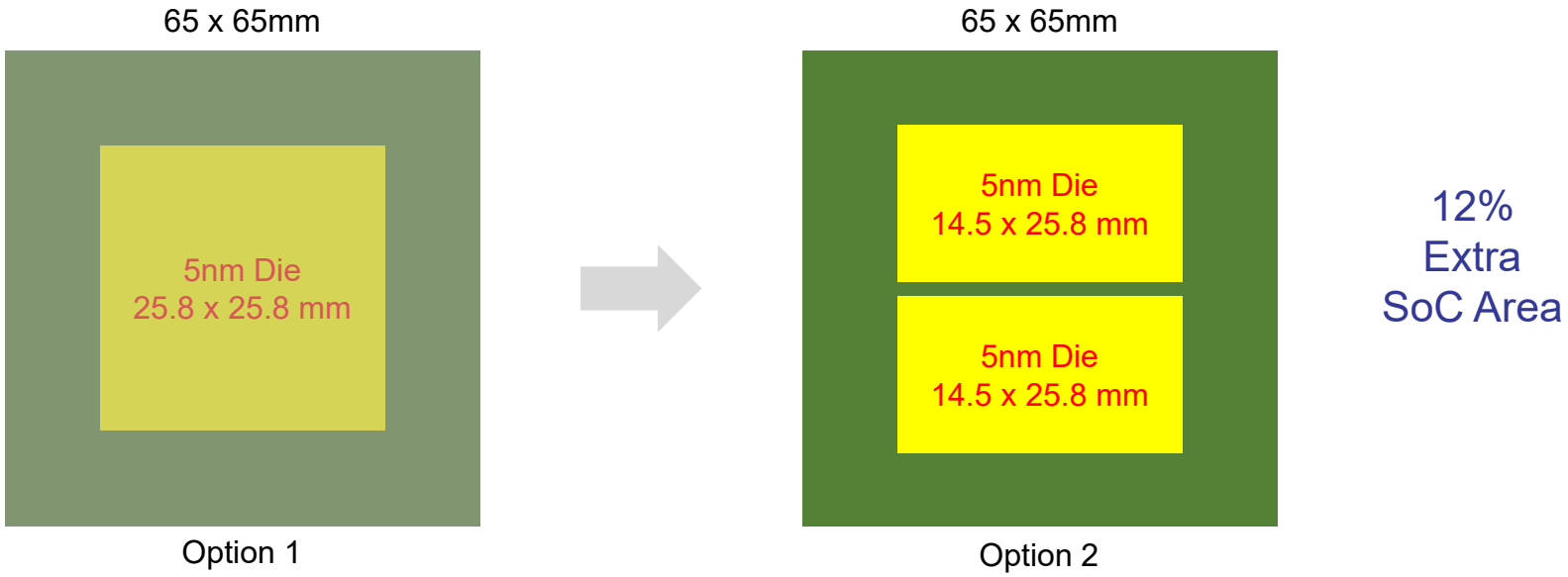
Test Case Examples

Test Cases



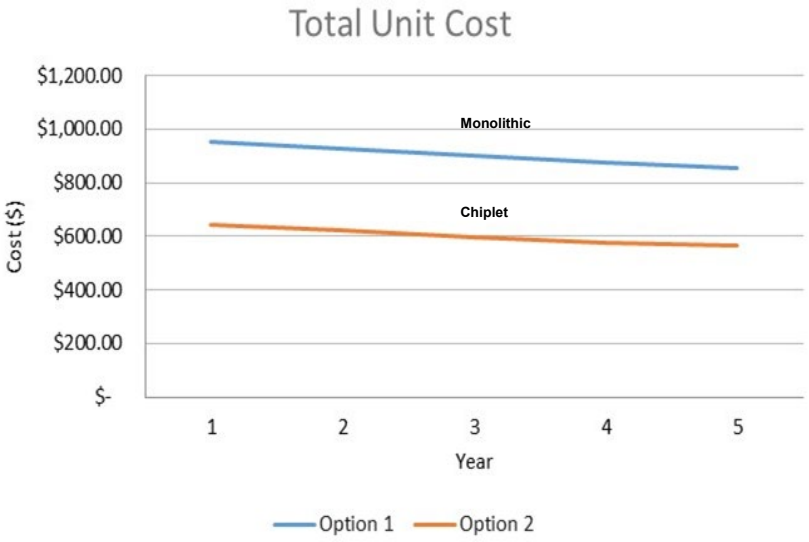
- Test Case 1: Monolithic vs. chiplet options for standard single die Flip Chip BGA
  - 12% extra silicon area
  - 32% less expensive
- Test Case 2: Integrating 16 chiplets onto a substrate
  - 12% extra silicon area
  - 20% less expensive
- Test Case 3: Large ASIC integrated with 4 HBM vs. ASIC divided into 2 chiplets integrated with 4HBMs
  - 8% extra silicon area
  - 30% less expensive

Scenario 1

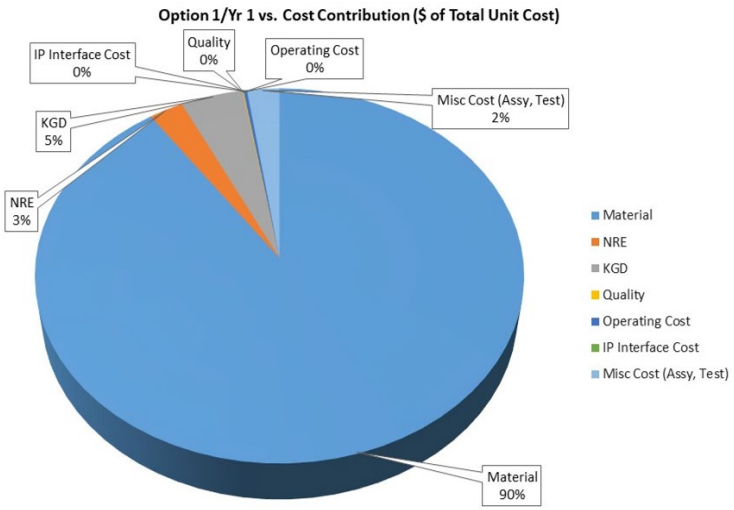


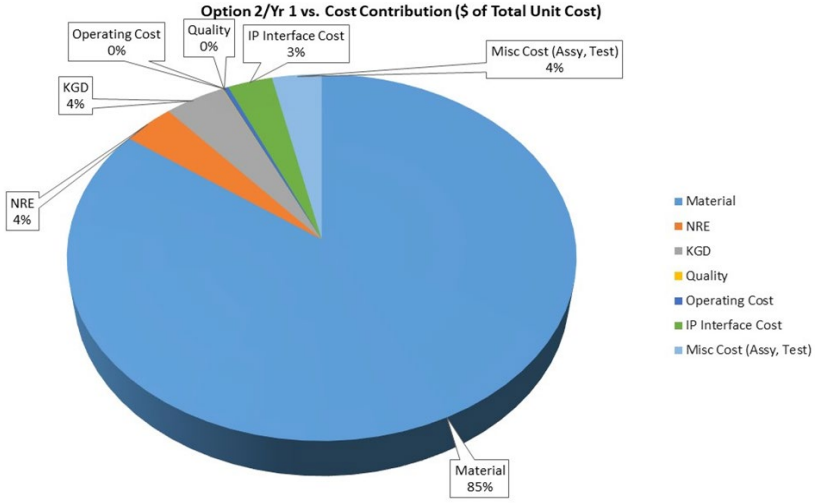
- Cost wise, should we split a monolithic die into two smaller ones?
- Several input parameters assumed in model
- No interposer

Scenario 1: Model Results

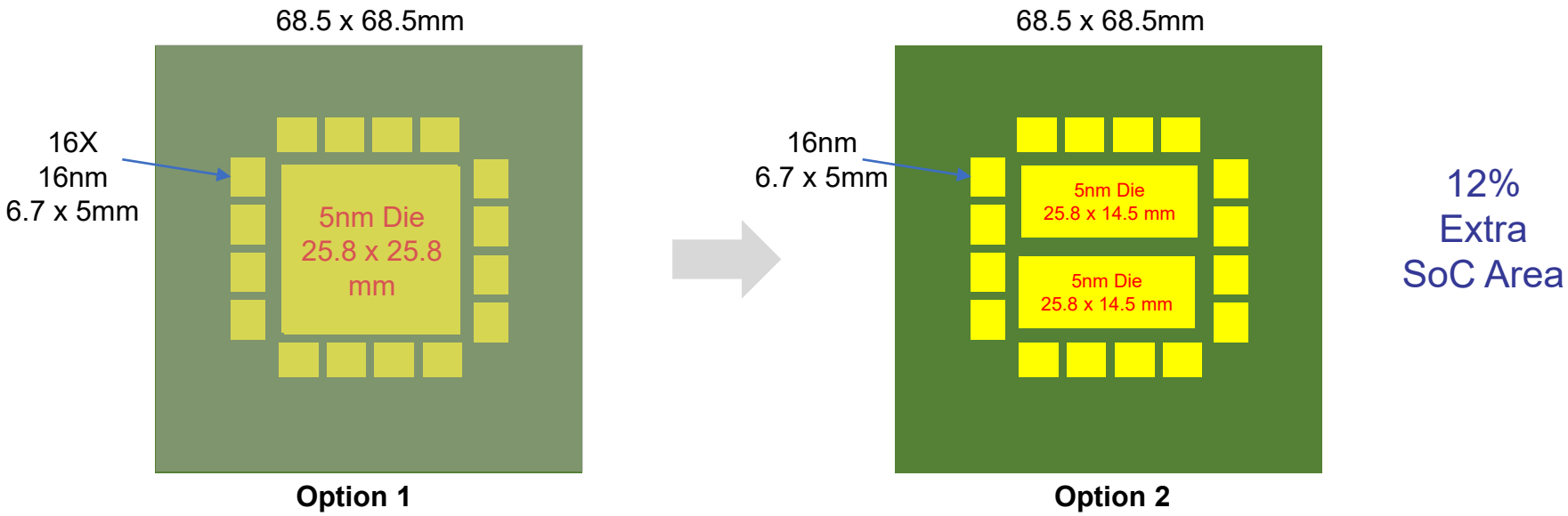


- Chiplets are 32% less expensive than monolithic
- Cost Drivers:
  - Material > KGD > NRE > Misc Cost > IP Interface > Operating Cost > Quality
- Chiplets in this case make economic sense.



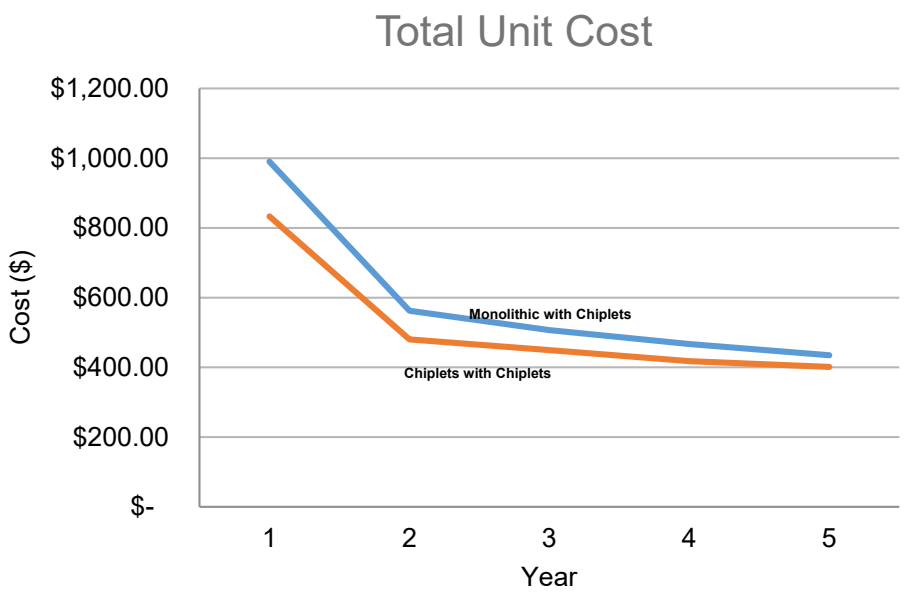


Scenario 2

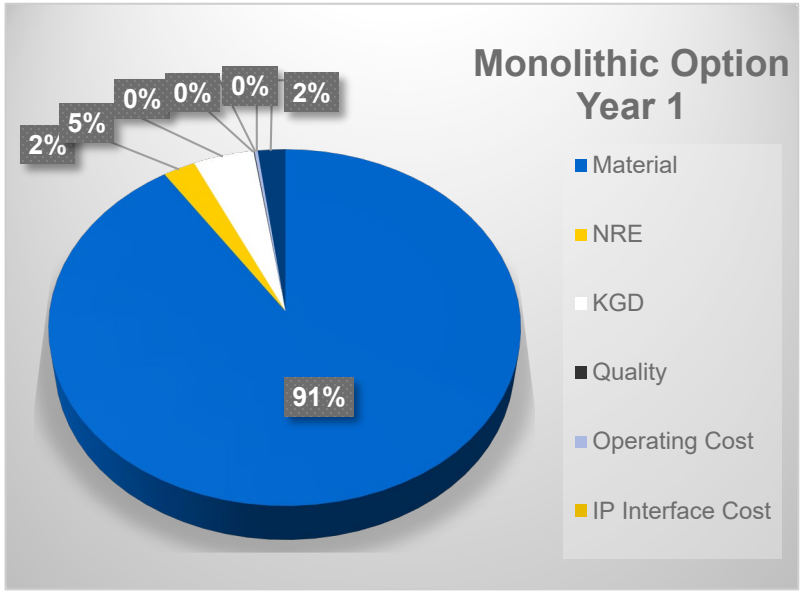


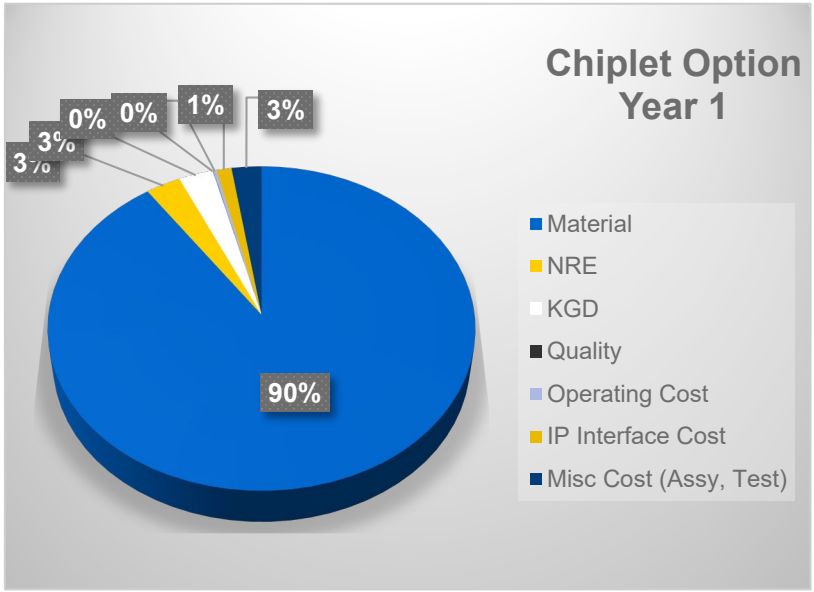
- Cost wise, should we split a monolithic die into two smaller ones?
- Several input parameters assumed in model
- No interposer
- Wafer yield goes down over time, volumes go up over time

Scenario 2: Model Results

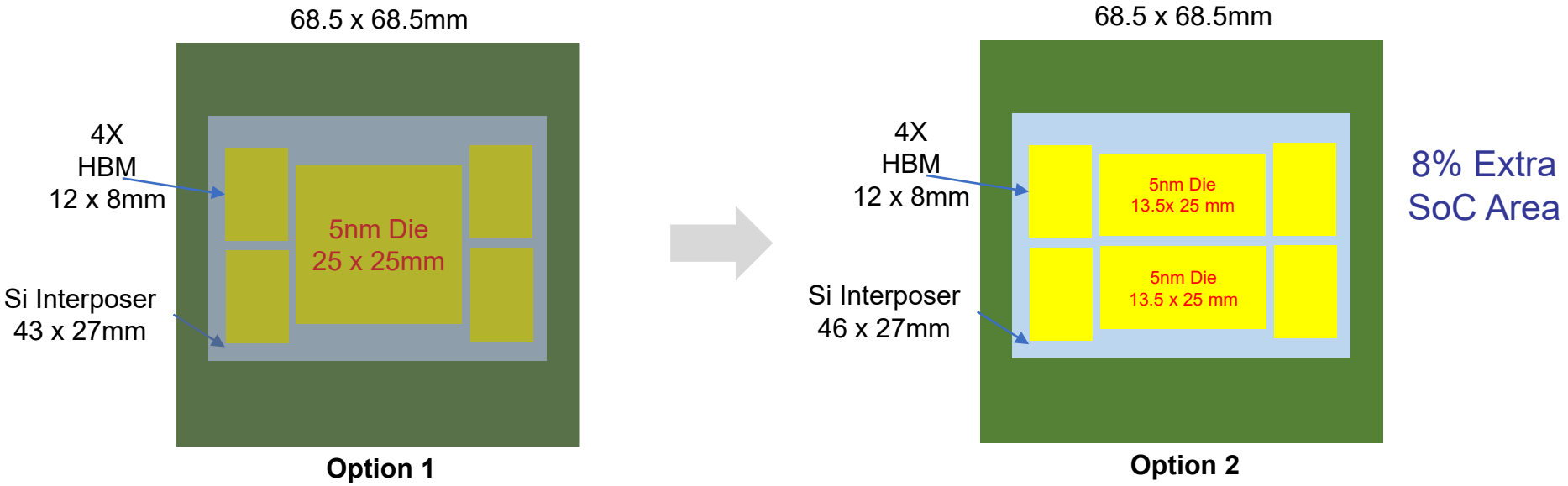


- Chiplets are **20% less expensive** than the monolithic version
- Cost Drivers:
  - Material > NRE > KGD > Misc. Cost > IP Interface > Operating Cost > Quality
- Main die split in this case makes economic sense.



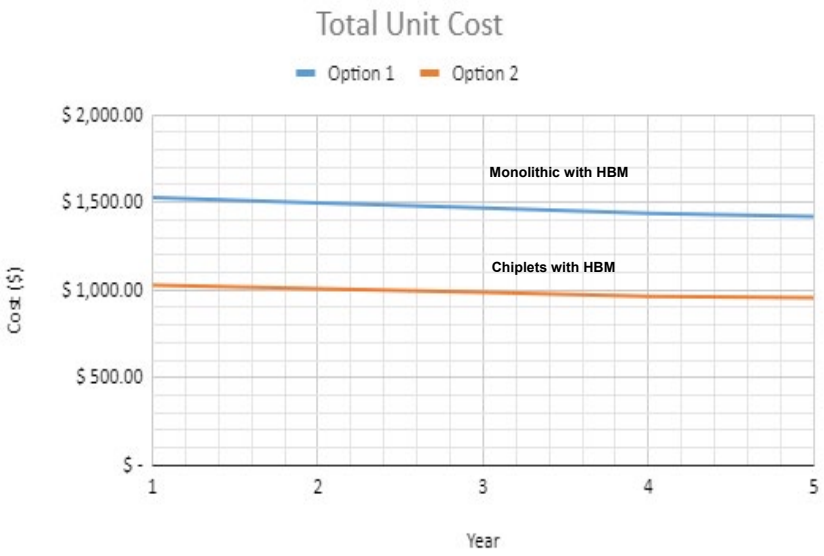


Scenario 3

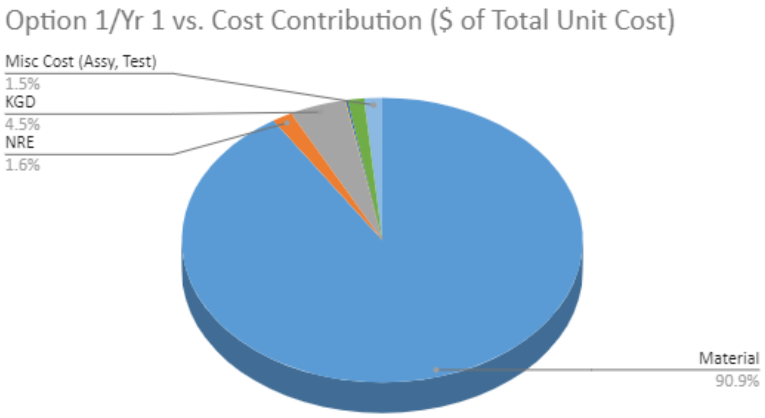


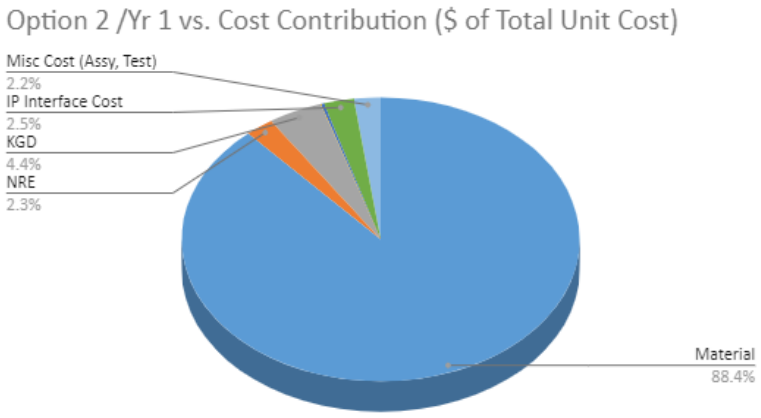
- Cost wise, should we split a monolithic die into two smaller ones?
- Several input parameters assumed in model

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.





Google

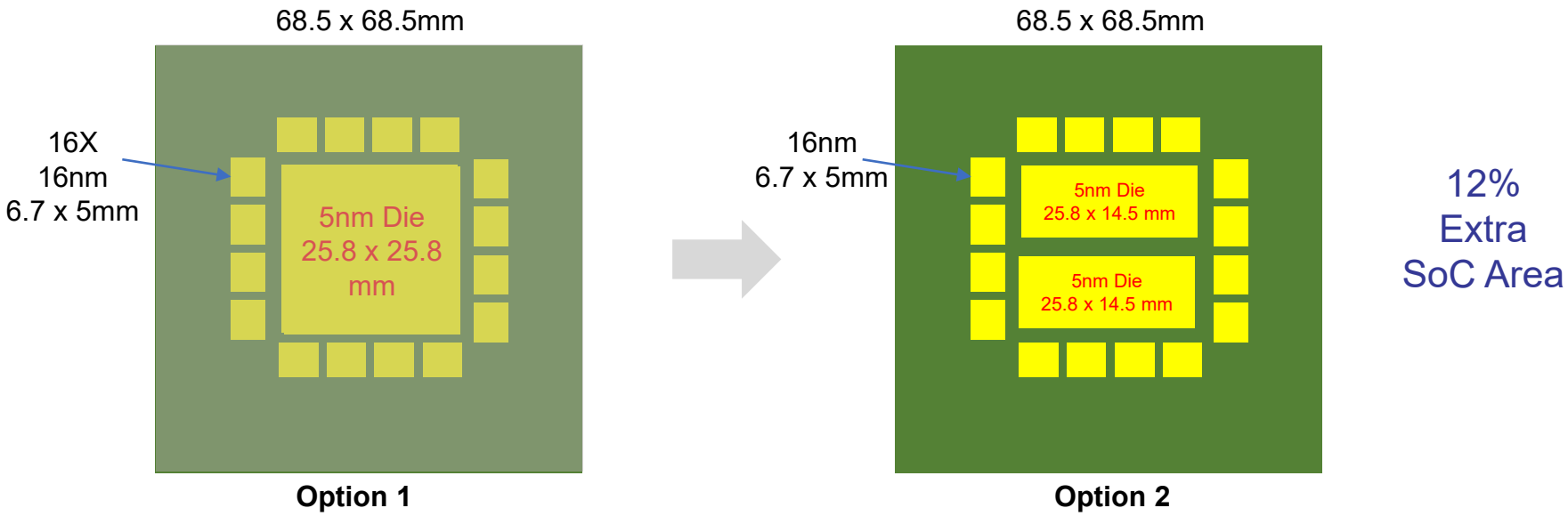
Effect of Uncertainty

Future is always uncertain

Google

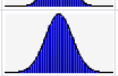
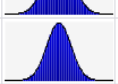
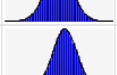
- The previous cases assume that all the factors are known and fixed
- In real life, several factors are uncertain and could vary wildly. For example:
  - Wafer Yield could change differently over time
  - Forecast Demand / shipment volumes can also change significantly over time
  - Assembly Yield also 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

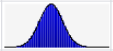
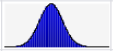
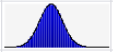
Scenario 2 with Uncertainty



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

Scenario 2 with Uncertainty: Inputs

Name	Graph	Minimum	Maximum	Mean	Mode	Median	Std. Deviation
Data - Option 1							
Option1 Assembly Yield		92.0000%	94.9997%	93.5000%	94.9850%	93.4999%	0.8661%
Option1 Yr1 FD		23,484.00	175,301.43	99,999.75	98,745.63	99,995.53	19,998.71
Option1 Yr2 FD		36,811.80	266,618.41	150,000.48	149,624.26	149,994.39	29,998.55
Option1 Yr3 FD		49,963.40	460,962.55	250,001.28	249,372.96	249,990.51	50,010.95
Option1 Yr4 FD		247,442.16	1,744,300.02	1,000,000.55	992,477.63	999,997.27	199,983.49
Option1 Yr5 FD		130,253.59	1,768,447.10	999,990.31	1,007,522.37	999,957.04	200,027.96
Option1 Yr1 DD		0.061241	0.137409	0.100000	0.100125	0.100000	0.010000
Option1 Yr2 DD		0.054872	0.125138	0.090000	0.089661	0.090000	0.009000
Option1 Yr3 DD		0.043892	0.105901	0.075000	0.074341	0.075000	0.007502
Option1 Yr4 DD		0.036956	0.082934	0.060000	0.059925	0.059999	0.006000
Option1 Yr5 DD		0.029626	0.068866	0.050000	0.050439	0.049999	0.005000

Option2 Assembly Yield		90.0003%	94.9997%	92.5000%	94.5250%	92.4997%	1.4434%
Option2 Yr1 FD		24,511.26	182,617.54	100,000.48	100,250.68	99,997.35	20,000.68
Option2 Yr2 FD		35,150.11	266,014.56	150,000.21	150,376.37	149,992.68	29,998.94
Option2 Yr3 FD		62,647.82	467,195.64	250,003.45	251,880.87	249,990.39	50,006.37
Option2 Yr4 FD		246,400.55	1,775,798.39	999,999.99	997,493.00	999,952.65	199,988.24
Option2 Yr5 FD		240,512.72	1,814,292.11	1,000,004.39	992,477.22	999,963.51	200,024.49
Option2 Yr1 DD		0.059526	0.137290	0.100000	0.100125	0.099998	0.010001
Option2 Yr2 DD		0.054297	0.123664	0.090000	0.090113	0.090000	0.009000
Option2 Yr3 DD		0.045924	0.103702	0.075000	0.075094	0.074998	0.007500
Option2 Yr4 DD		0.033357	0.083526	0.060000	0.059774	0.059999	0.006002
Option2 Yr5 DD		0.030819	0.069387	0.050000	0.050188	0.049999	0.005000
Option2 D2 Yr1 DD		0.060140	0.140262	0.100000	0.099624	0.099999	0.010000
Option2 D2 Yr2 DD		0.054563	0.123713	0.090000	0.090564	0.089999	0.008999
Option2 D2 Yr3 DD		0.046709	0.104765	0.075000	0.075470	0.074998	0.007500
Option2 D2 Yr4 DD		0.036837	0.083474	0.060000	0.060075	0.059999	0.006000
Option2 D2 Yr5 DD		0.030825	0.069978	0.050000	0.050063	0.050000	0.005000

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

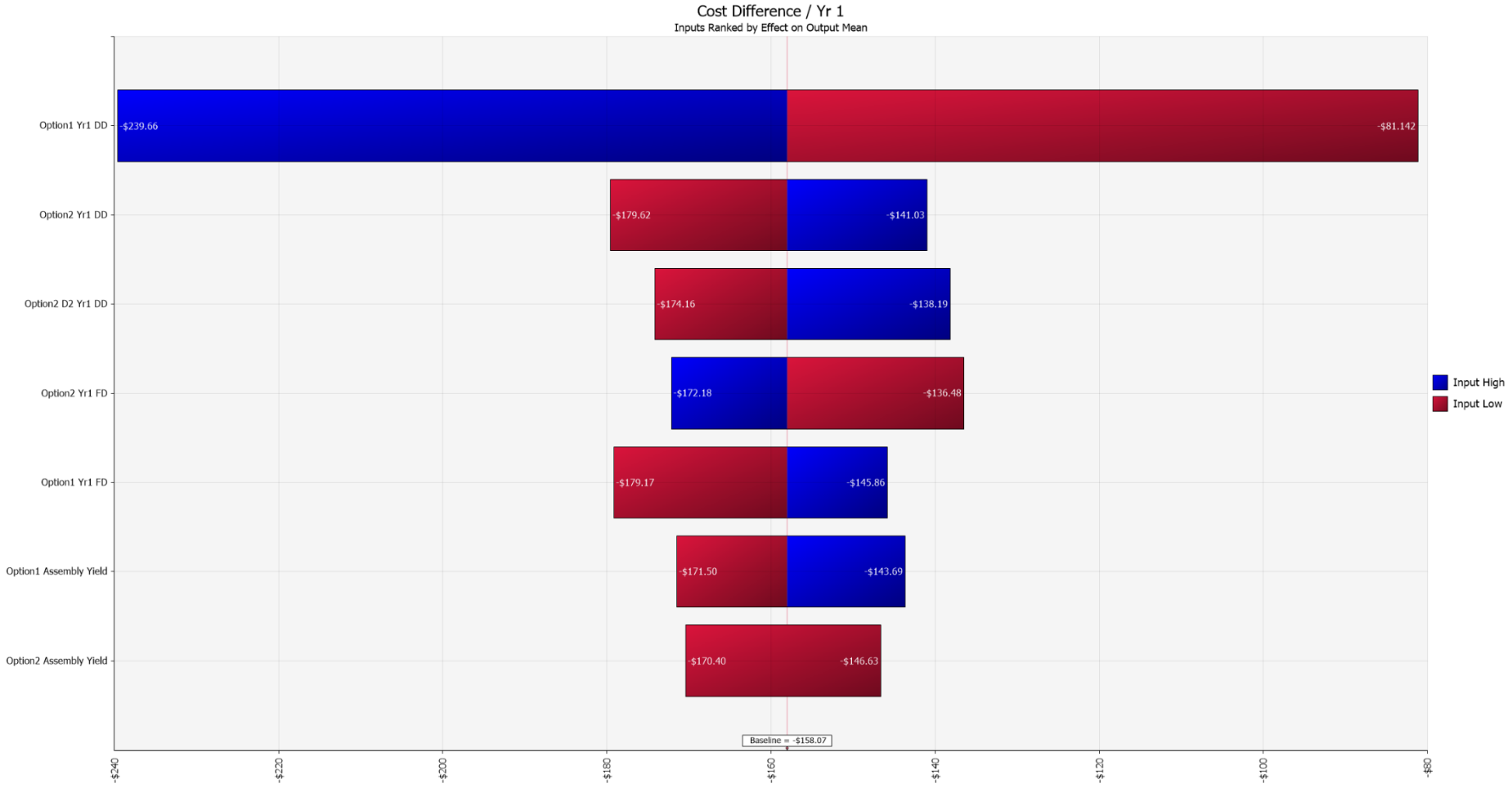
Scenario 2 with Uncertainty: Results

Cost Benefit	Year 1	Year 2	Year 3	Year 4	Year 5
Worst Case (5% Probability)	$10	$55	$36	$31	$19
Average (50% Probability)	$156	$82	$57	$48	$33
Best Case (5% Probability	$325	$110	$79	$64	$47

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

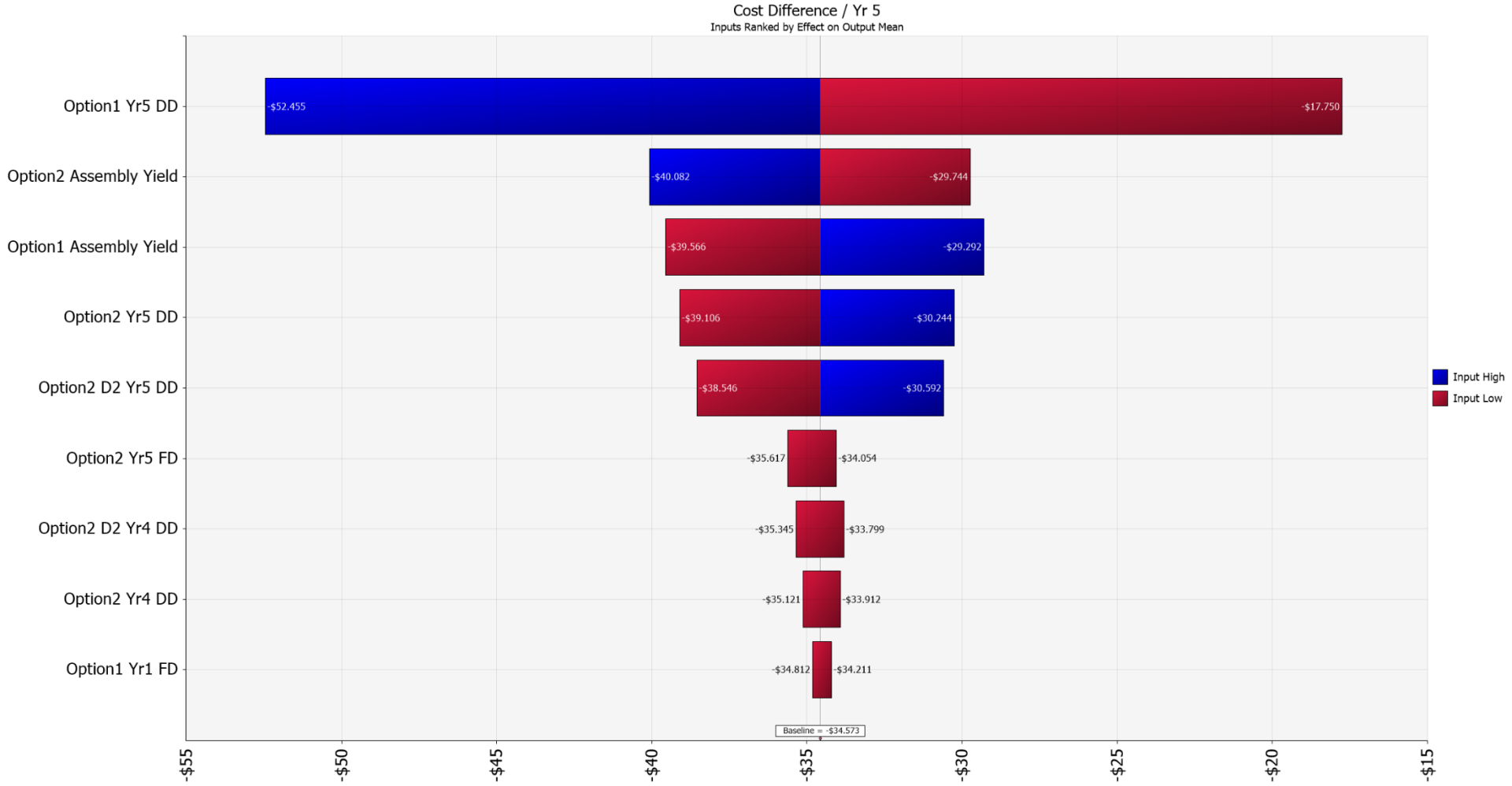
Scenario 2 with Uncertainty: Drivers



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]

Google

$$GDPW_i = \left\lfloor (Wfr_i - 6)\pi \left(\frac{Wfr_i}{4Eff A_i} - \frac{1}{\sqrt{2Eff A_i}} \right) \right\rfloor$$ Gross Die per Wafer

$$Pwafer_{t,i} = Pwafer_{t=t-1,i} \times \left(1 - \frac{dr_i}{100}\right)$$ Price per Wafer (t > 1)

$$Yield_{Murphy_{t,i}} = \left[\frac{1 - e^{-dd_{t,i} Eff A_i \times 0.01}}{dd_{t,i} Eff A_i \times 0.01} \right]^2 + REC_{yield}$$ Yield of each wafer I at time t (Murphy)

$$FUP_{Murphy_{t,i}} = \left[\frac{Pwafer_{t,i}}{INT(\text{GDPWi} \times Yield_{Murphy_{t,i}})} \right] + Probe\ Cost$$ Forecast Unit Price per die i at time t, (Murphy)

$$MatCost_{Murphy_{t,i}} = FUP_{Murphy_{t,i}} \times FD_{t,i}$$ Material Cost of die i at time t (Murphy)

$$SubsUnitCost_t = \frac{Subs\_Unit\_Price_t}{Subs\_Yield_t}$$ Substrate Unit Cost at time t

$$SubsCost_t = SubsUnitCost_t \times Subs\_Demand_t$$ Substrate Total Cost at time t

$$Misc\_Cost_t = Package\_Assy\_Cost_t \times FD_{t,substrate}$$ Miscellaneous Cost at time t

$$Assy\_Yield_s = Assy\_Per\_Step\_Yield_s^{\sum_{i,s}^{I,S} Assy}$$ Yield at each assembly step

$$Assy\_Scrap_{Murphy_t} = \sum_{i,s}^{I,S} MatCost_{Murphy_{t,i,s}} \times Assy_{t,i,s} \times (1 - Assy\_Yield_s)$$ Assembly Scrap at time t (Murphy)

$$Total\_IP\_Cost_t = IP\_Cost_t + \left(\frac{IP\_Pct\_per\_ASP_t \times ASP_t}{100} \times FD_{t,i} \right)$$ IP interface Cost at time t

$$Field\_Quality\_Cost_{Murphy_{t,i}} = \frac{Field\_Quality\_FR_{i,t}}{1000000} \times FUP_{Murphy_{t,i}} \times FD_{t,i}$$ Field Quality Cost of die i at time t (Murphy)

$$WSt_i = (WS_a + WS_h) \times \left(\sum_i^{WSc_i} WSd_i + (1 - Yield_{ti} \times WSx_i) \times \sum_i^{WSr_i} WSd_i \right)$$ *Wafer Sort if Applicable*

$$FTt_i = (FT_a + FT_h) \times \left(\sum_i^{FTc_i} FTd_i + \left(1 - Yield_{ti} \times (WSx_i - FTx_i)\right) \times \sum_i^{FTr_i} FTd_i\right)$$ 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