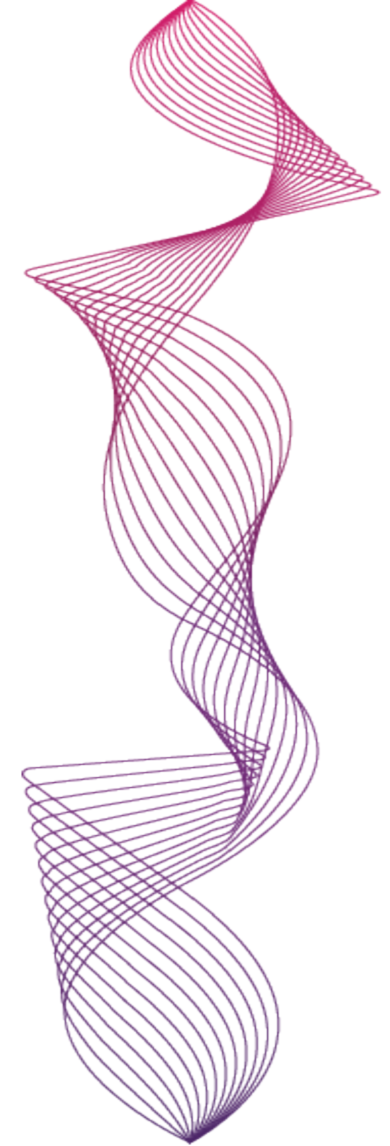


3D Design and Performance

DBI®-Enabled Next Generation SoC Architectures

Javi DeLaCruz

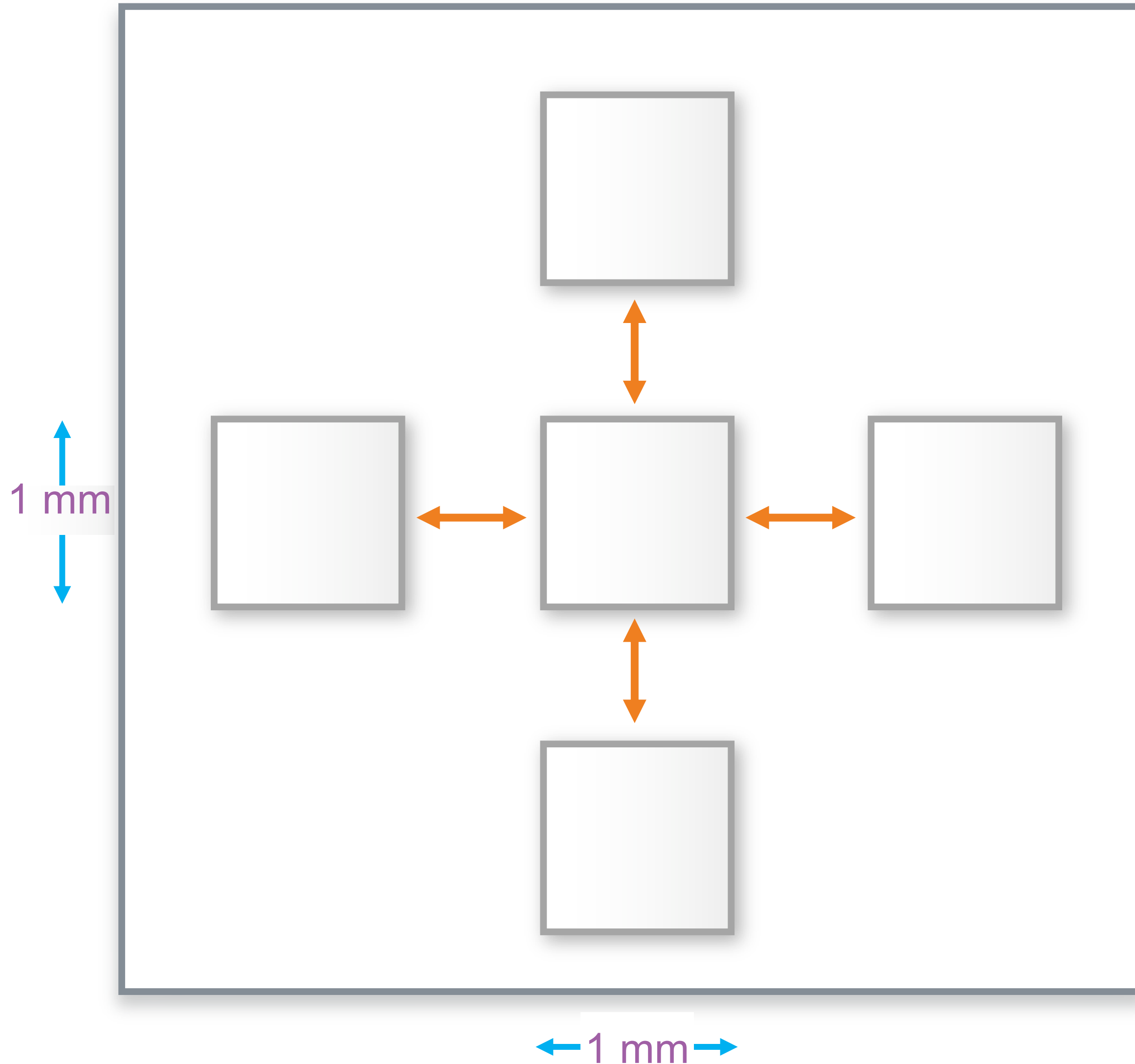
Special Thanks to for contributions

Adapted from S3S Conference 2019 presentation

Agenda

- Limitations of Computing
- Fine-pitch 3D interconnect confers unique, powerful, new capabilities
  - *Lower power, higher performance, reduced area*
- Production-proven 3D technologies: ZiBond® and DBI
- Design in 3D instead of stacking 2D designs
- Reticle Limitations Emerging
- Case study in High Performance Compute

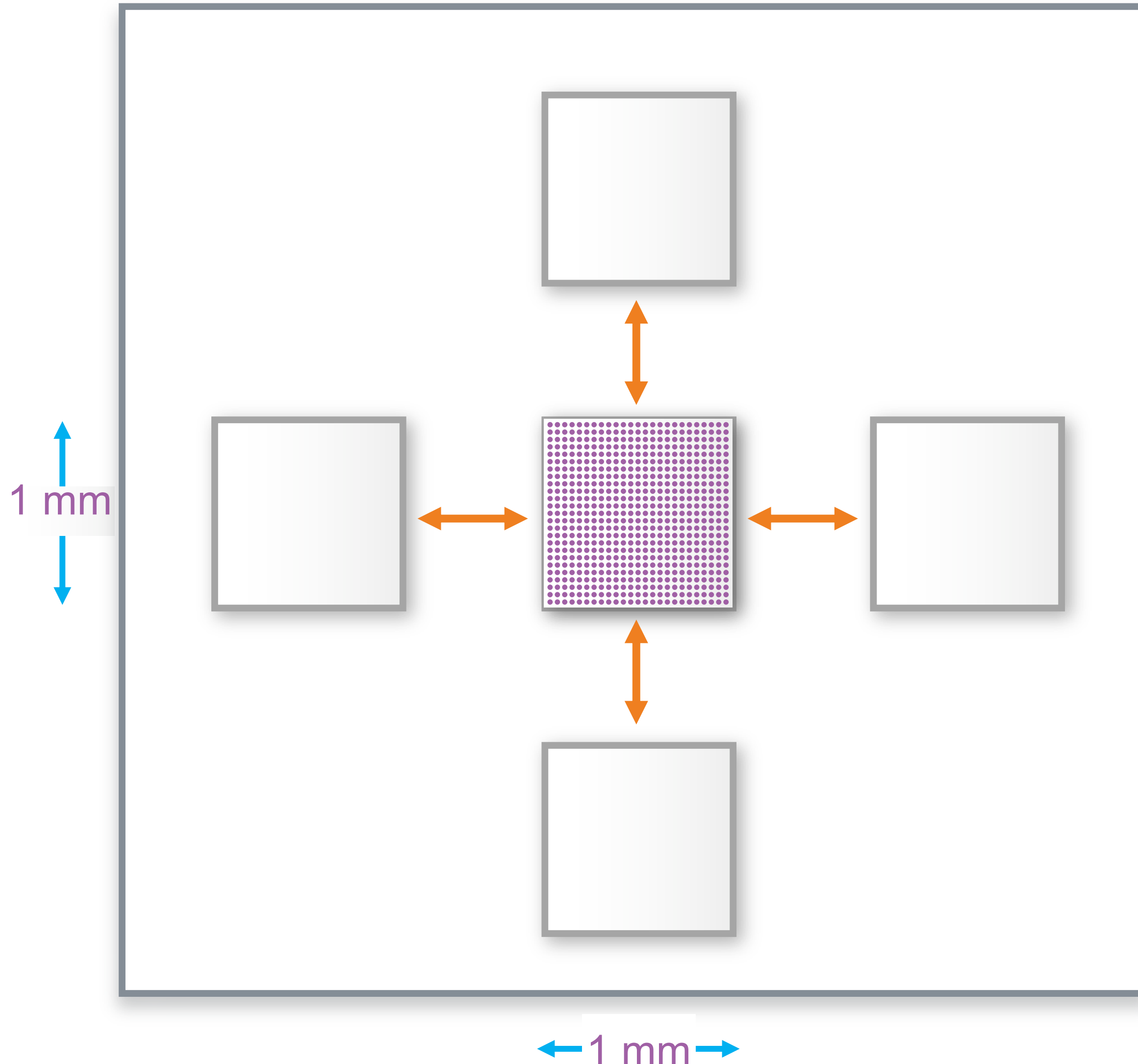
Main computation bottleneck is connectivity



With 10nm manufacturing…

- 12 signals/μm of beachfront on middle layers
- 4 middle layers **~100,000 connections / mm^2**

Main computation bottleneck is connectivity



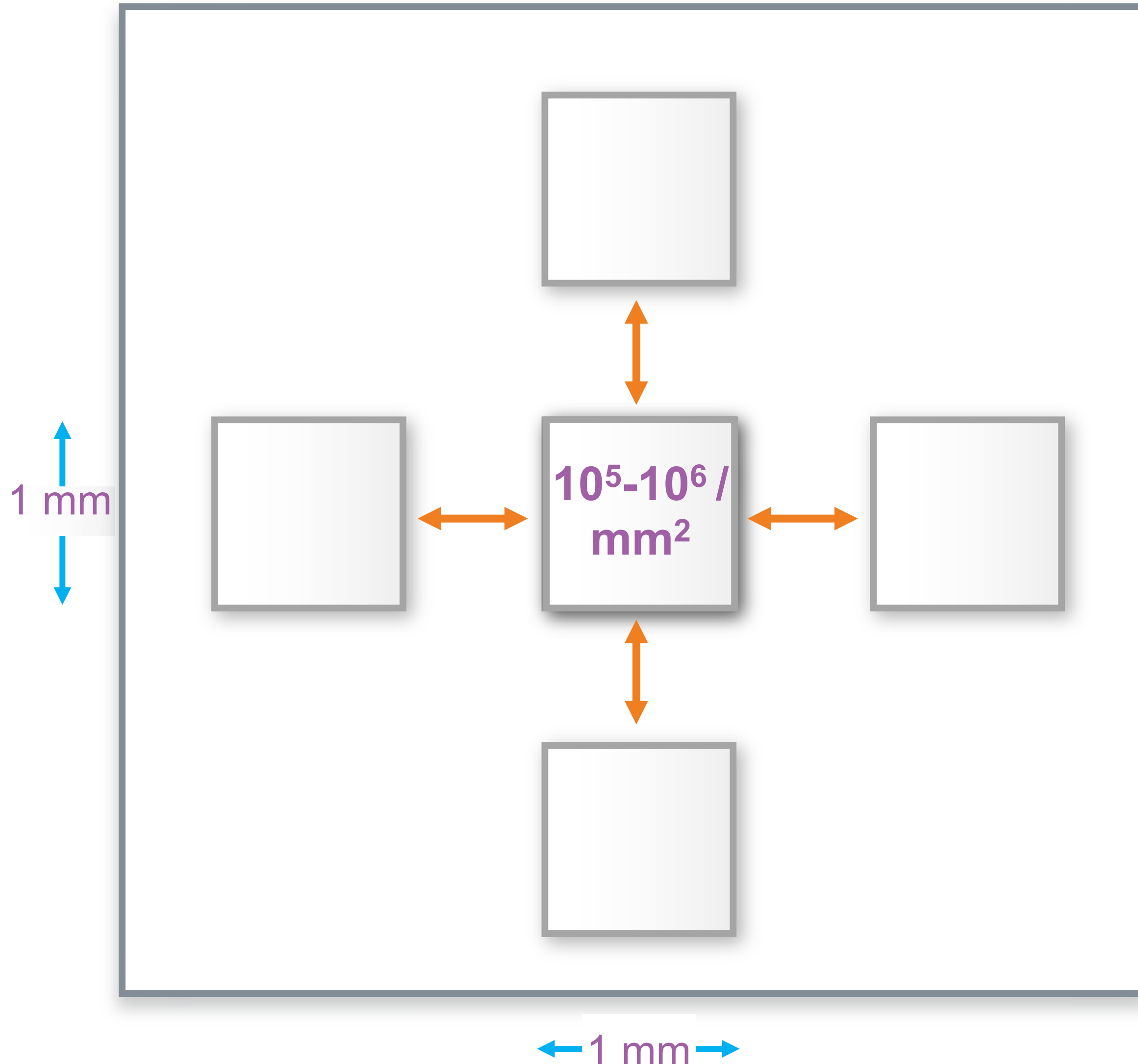
With 10nm manufacturing…

- 12 signals/µm of beachfront on middle layers
- 4 middle layers **~100,000 connections / mm²**

With most advanced TSVs…

- Only **625 connections / mm²**

Main computation bottleneck is connectivity



With 10nm manufacturing…

- 12 signals/μm of beachfront on middle layers
- 4 middle layers **~100,000 connections / mm^2**

With most advanced TSVs…

- Only **625 connections / mm^2**

Interface Between Die

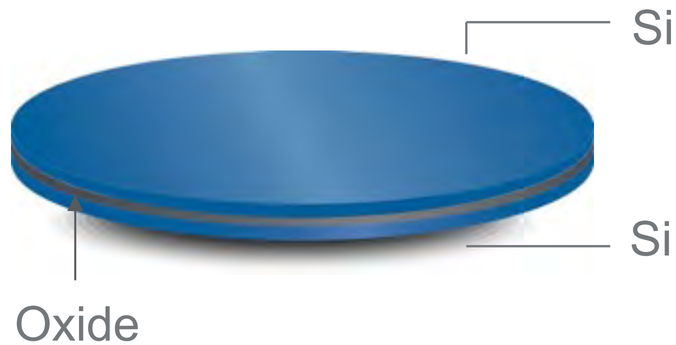
- **What's the best interface for 2.5D and 3D?   …the answers may be different**
- **Adding standard interfaces reduces the benefit of 3D design**
- **Leverage smaller load between die than within die**
- **Internal interconnects across die layers (AXI, APB, ASB, NoC, SRAM Bus)**
- **Folding alone, without planning improves average net length by 30%**
- **Deliberate 3D architectural planning can shrink routes from mm to µm**

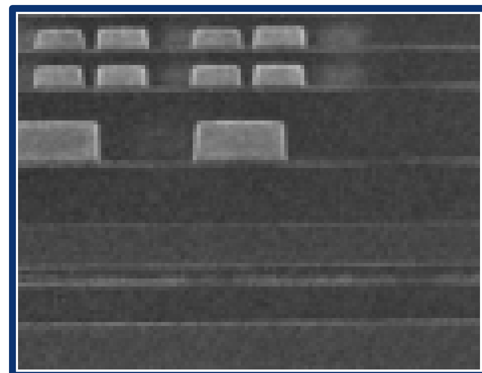
Interface between die can be the same as (or better than) interfaces within die

ZiBond & DBI 3D wafer/die bonding solutions

ZiBond

Direct Bonding

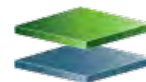




Courtesy Chipworks/Sony

Wafer to Wafer (W2W) Bonding

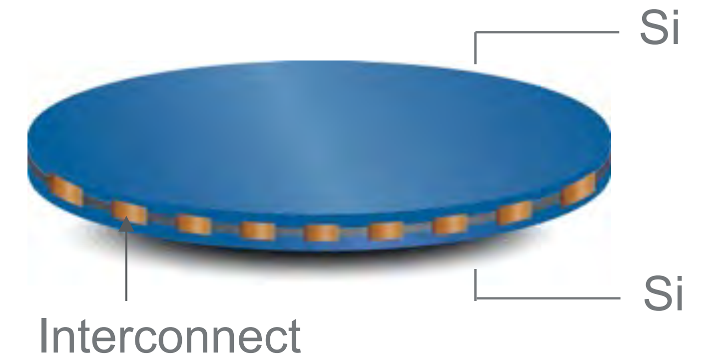
Die to Wafer (D2W) Bonding

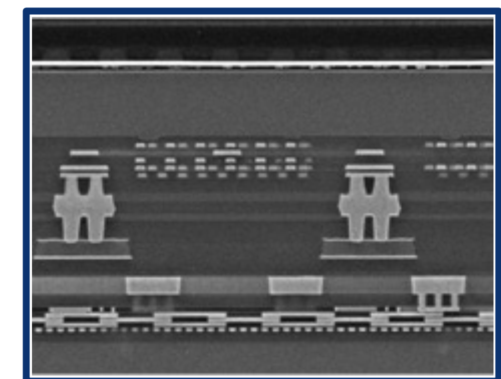


Die to Die (D2D) Bonding

DBI

Hybrid Bonding





Courtesy Chipworks/Sony

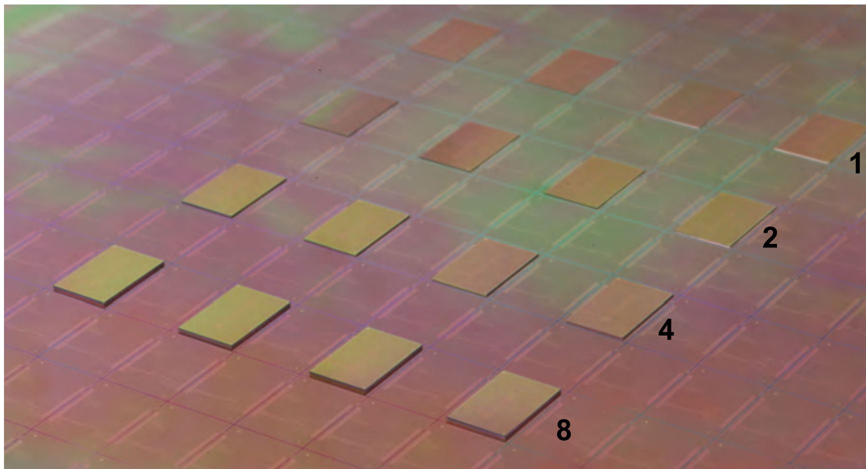
XPERI

Semiconductor Technologies | invensas

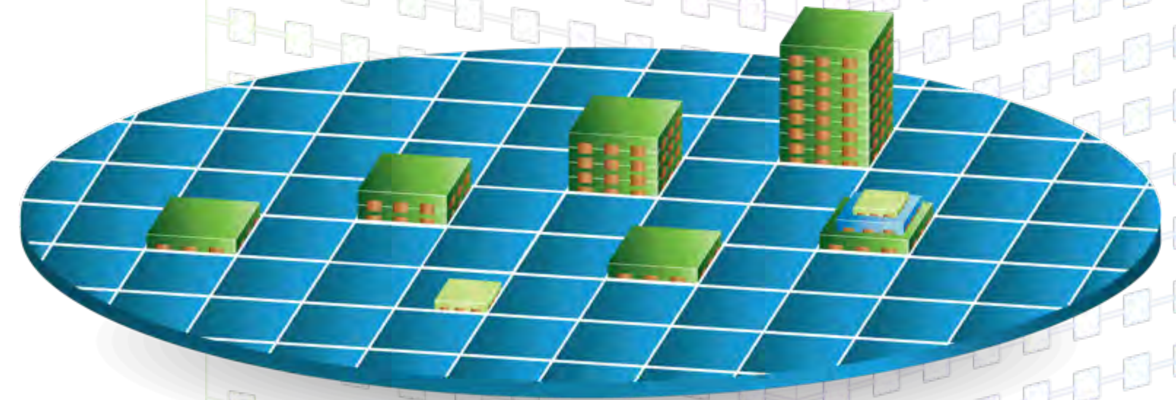
DBI® Ultra

Die to Wafer Hybrid Bonding

The Ultimate 2.5D and 3D Integration Technology for High-Performance Computing

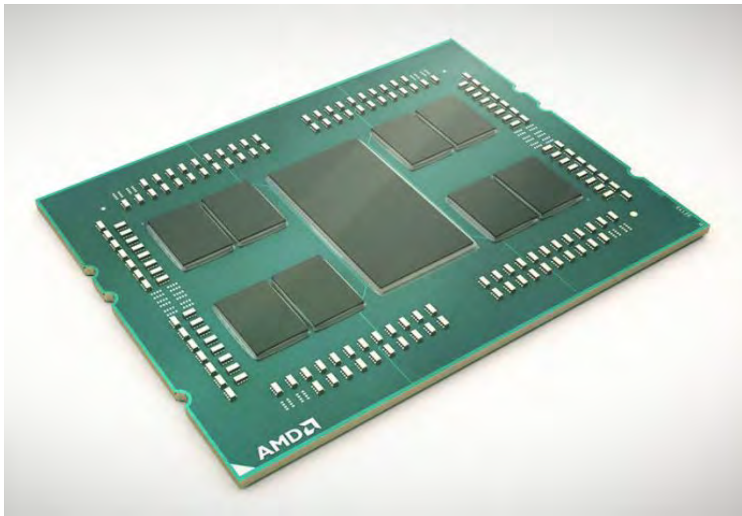


DBI Ultra Image: Gao et al; ECTC 2019

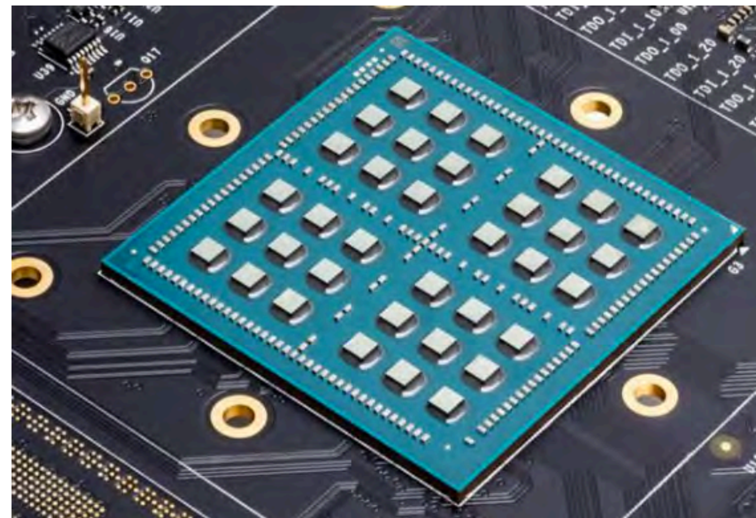


Reticle Buster Problem

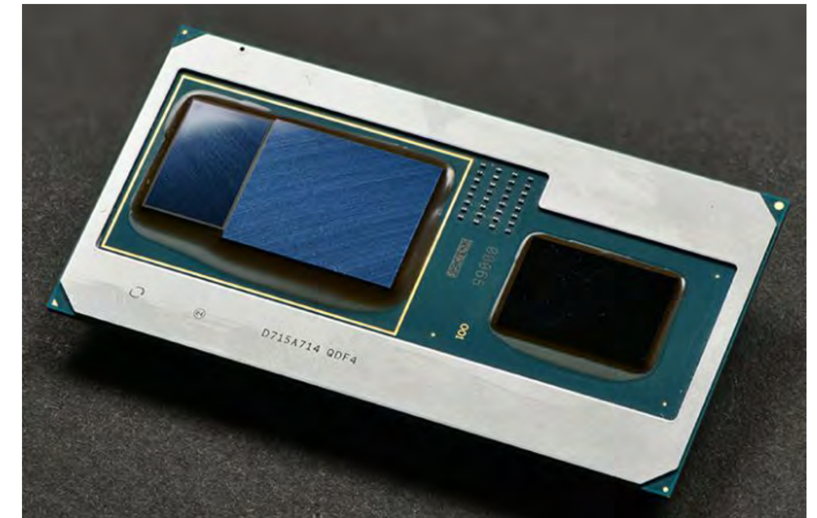
- The Industry is reaching a high hurdle with the reticle limits
- Impacts on yield, performance, cost, etc.
- Several ways to address this, which include chiplets



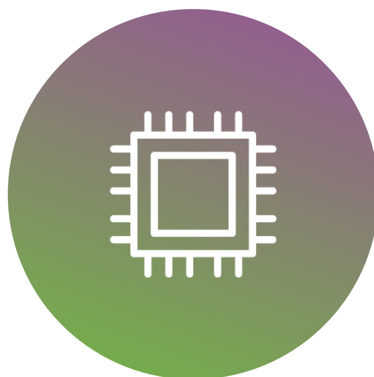
AMD EPYC 2 Rome
Image from www.servethehome.com



NVIDIA Deep Neural Network Accelerator
Image from HotChips 2019, Krizhevsky et al.



Intel 8th Generation Core with Radeon RX Vega M Graphics
Image from Anandtech



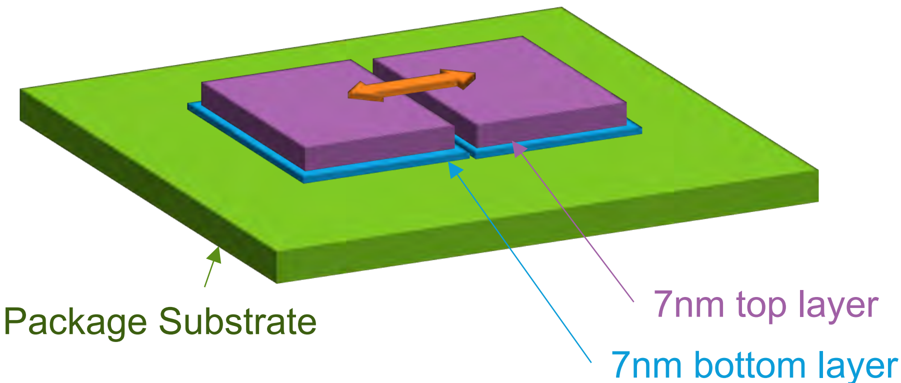
51.2Tbps Switch

High Performance Compute Case Study

High Performance Compute Analysis

- 51.2 Tbps Switch requires ~4 reticles at 7nm
- 512 lanes of 112Gbps SerDes off package
- Same logic/memory area in each solution, DBI Ultra
- Logic and memory on both layers when stacked.
  IO on top die due to SerDes hard IP

2.1 or 2.5 Interconnect	2 Stacks of 2 Die	2.5D Array of 4 Die
USR (no interposer)	Option A	Option C
HBI (Stitched interposer)	Option B	Option D
Native		Option E

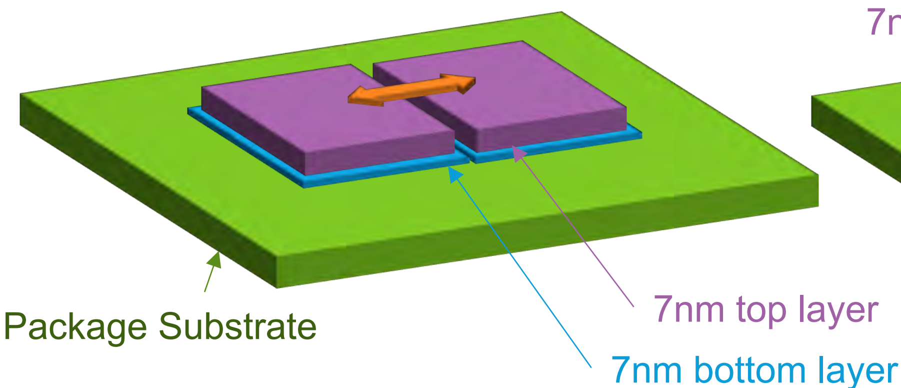


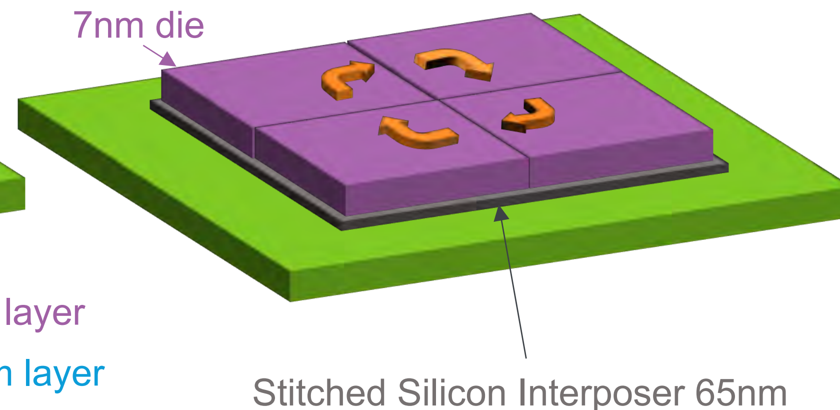
Option A
Option B includes interposer

High Performance Compute Analysis

- 51.2 Tbps Switch requires ~4 reticles at 7nm
- 512 lanes of 112Gbps SerDes off package
- Same logic/memory area in each solution, DBI Ultra
- Logic and memory on both layers when stacked.
  IO on top die due to SerDes hard IP

2.1 or 2.5 Interconnect	2 Stacks of 2 Die	2.5D Array of 4 Die
USR (no interposer)	Option A	Option C
HBI (Stitched interposer)	Option B	Option D
Native		Option E





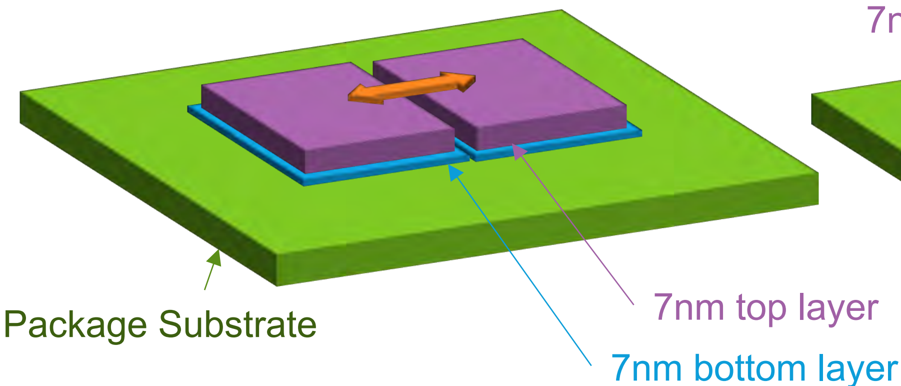
Option A
Option B includes interposer

Option C has no interposer
Option D includes interposer

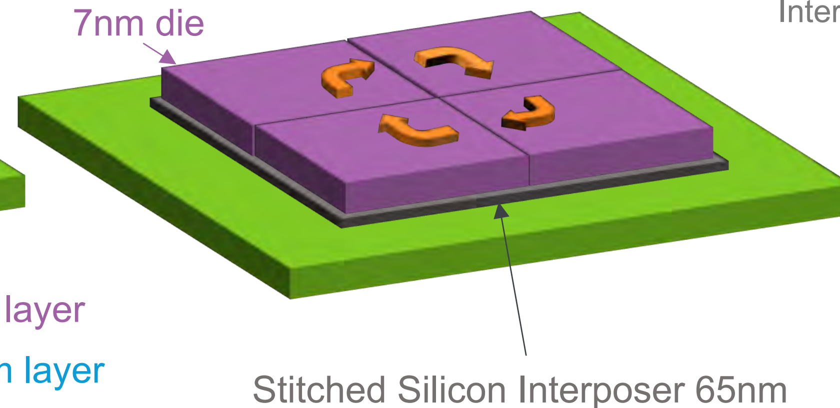
High Performance Compute Analysis

- 51.2 Tbps Switch requires ~4 reticles at 7nm
- 512 lanes of 112Gbps SerDes off package
- Same logic/memory area in each solution, DBI Ultra
- Logic and memory on both layers when stacked. IO on top die due to SerDes hard IP

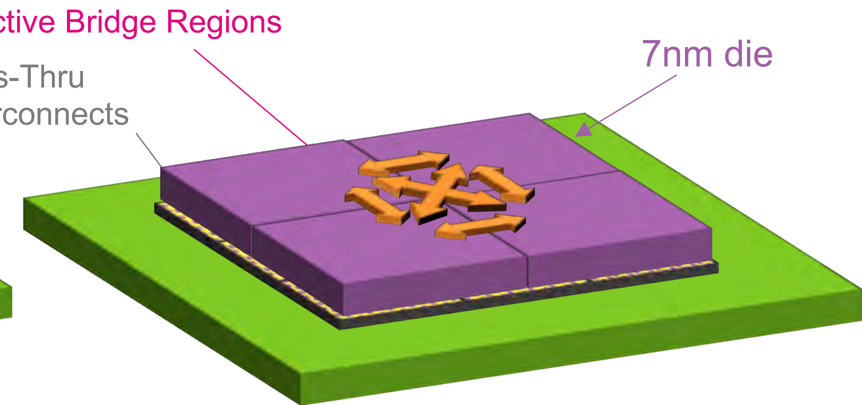
2.1 or 2.5 Interconnect	2 Stacks of 2 Die	2.5D Array of 4 Die
USR (no interposer)	Option A	Option C
HBI (Stitched interposer)	Option B	Option D
Native		Option E



Option A
Option B includes interposer



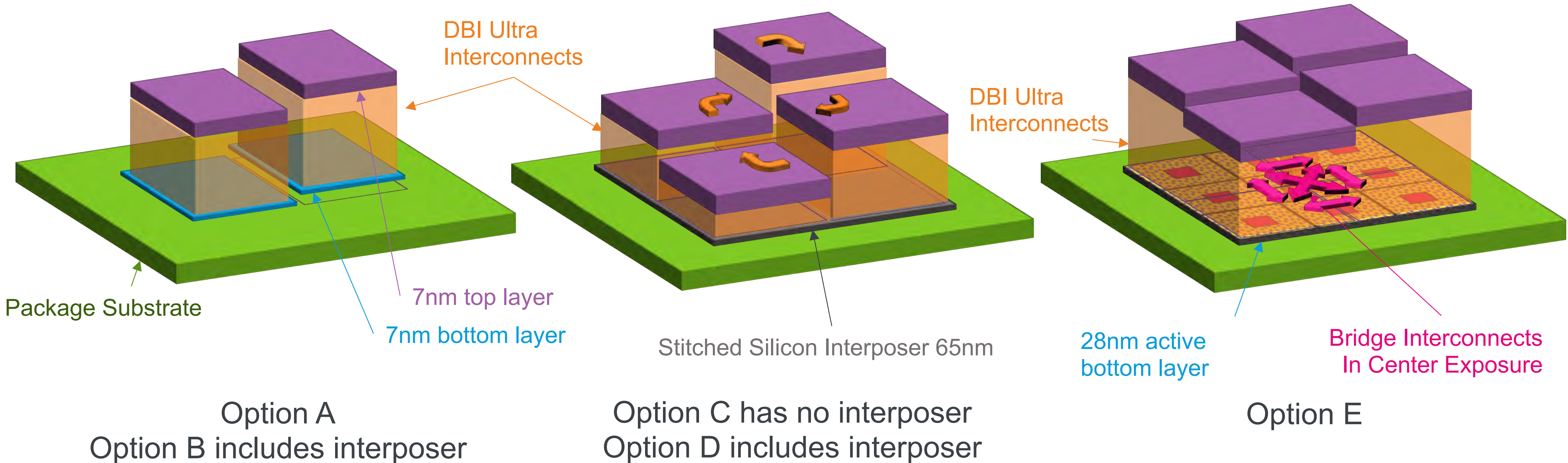
Option C has no interposer
Option D includes interposer



Option E

Obstacles and Advantages in Analysis

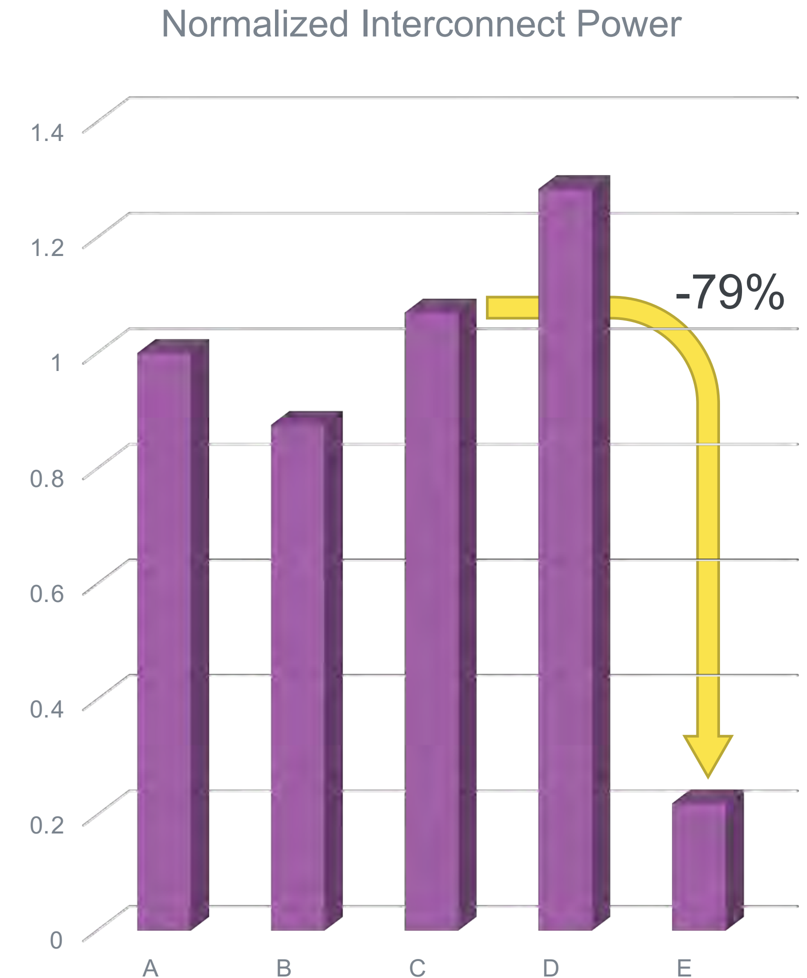
- Utilizing DBI Ultra for yield improvement
- Unable to floorplan the USR in Option A due to limited beachfront with two rows of USR.
- Option E utilizes active and unstitched large base die in 28nm

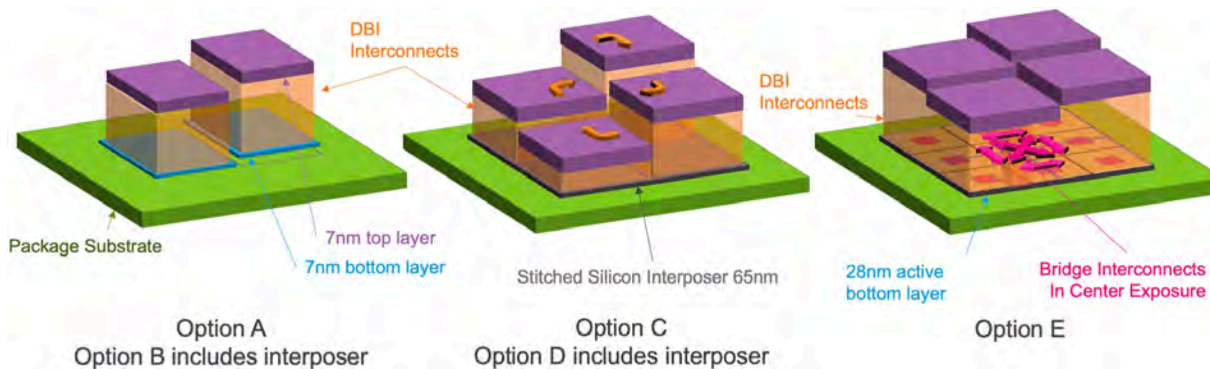


Comparative Power Analysis

Interface	2.1D + 3D 2 Stacks of 2 Die	2.5D Array of 4 Die
USR	Option A	Option C
HBI	Option B	Option D
Native		Option E

- Only the lateral chip-chip interconnect power considered
- **Native interconnects** on Option E consume the least power



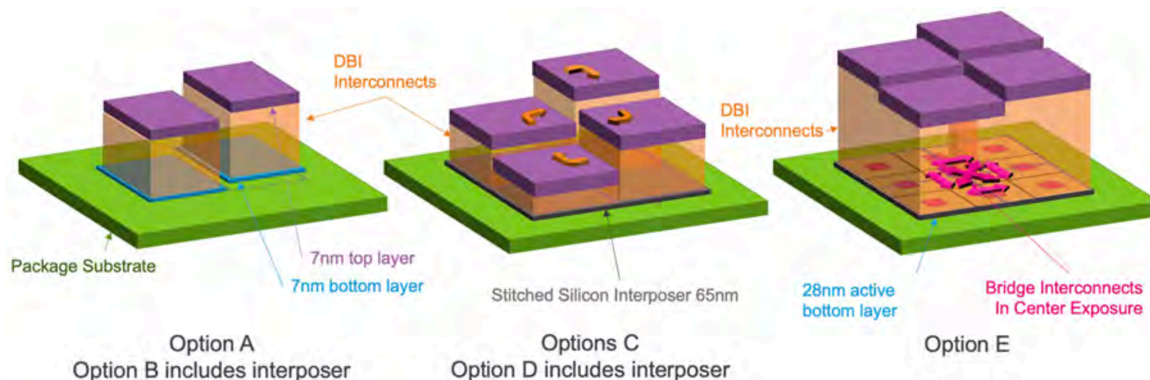


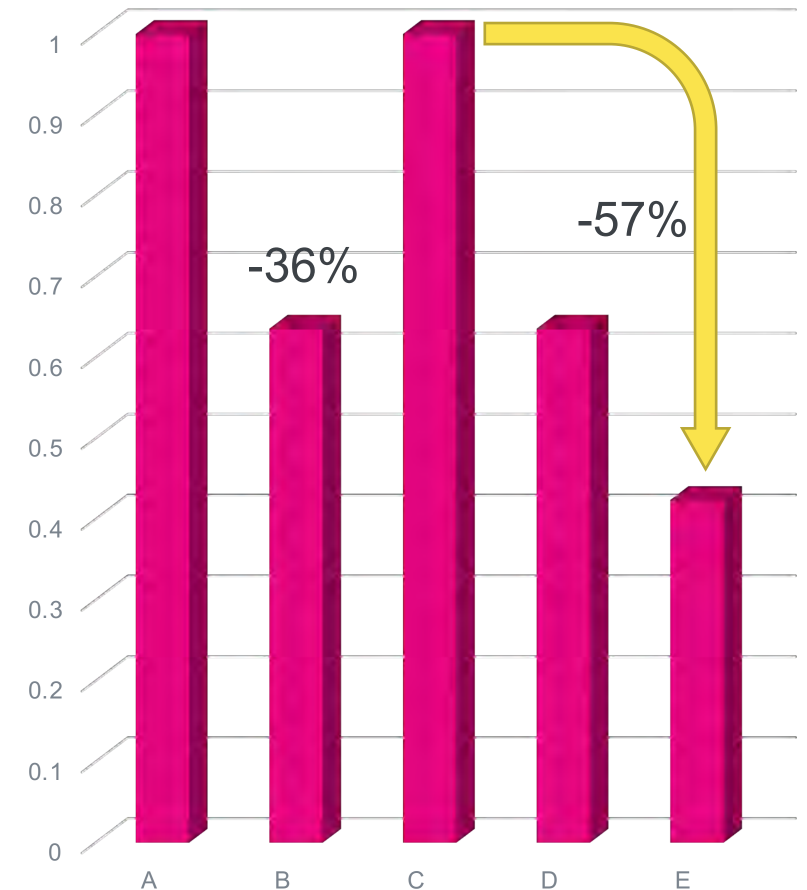
XPERI

Comparative Latency Analysis

Interface	2.1D + 3D 2 Stacks of 2 Die	2.5D Array of 4 Die
USR	Option A	Option C
HBI	Option B	Option D
Native		Option E

- HBI has an inherently lower latency than a USR interface
- Native interconnects have a 57% improvement over using a USR SerDes



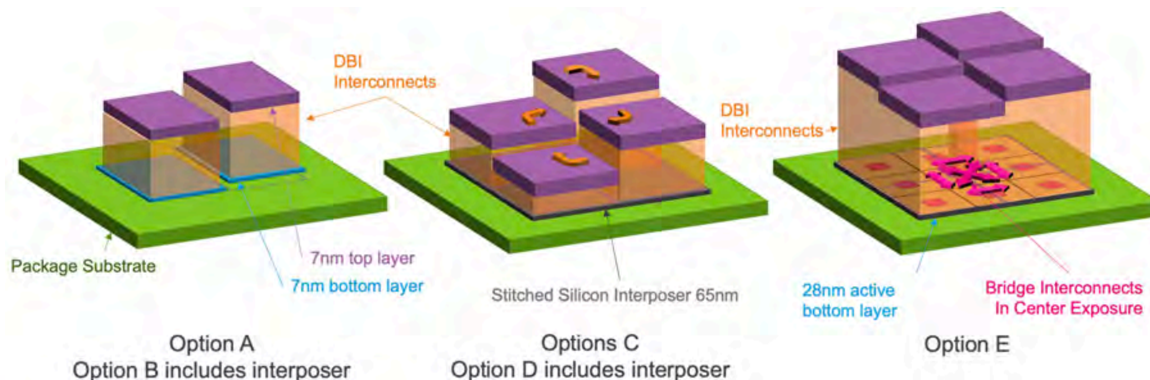


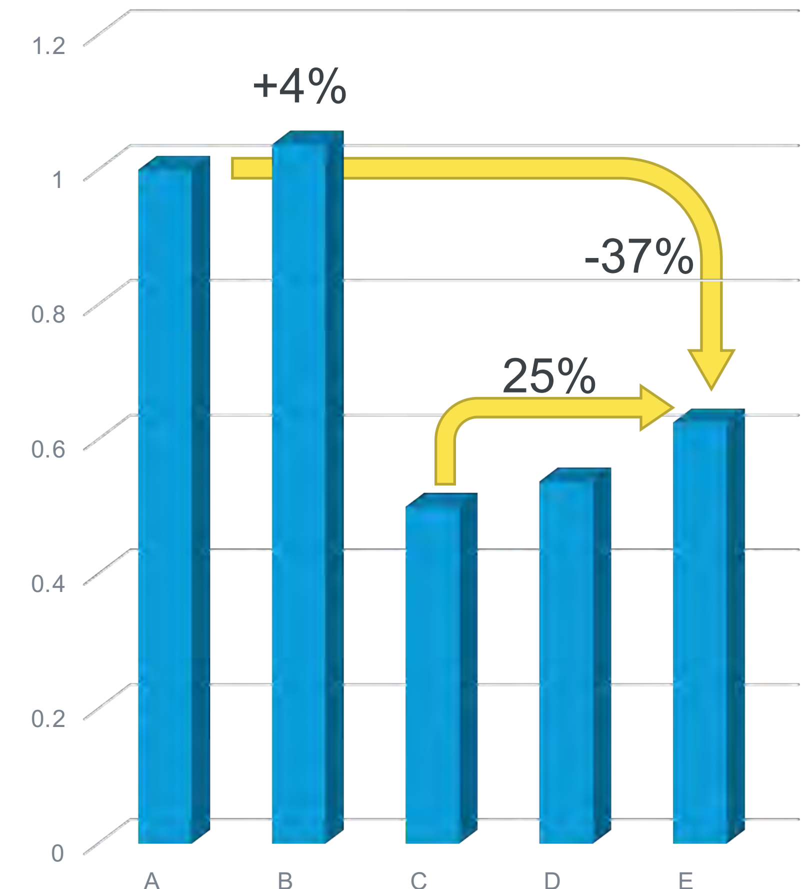
XPERI

Comparative Mask NRE Analysis

Interface	2.1D + 3D 2 Stacks of 2 Die	2.5D Array of 4 Die
USR	Option A	Option C
HBI	Option B	Option D
Native		Option E

- Options A and B comprise two 7nm tapeouts
- Option B had higher NRE due to additional cost of 65nm interposer
- Option C is the simplest with a single 7nm tapeout
- Option E has only one 7nm and one 28nm tapeout

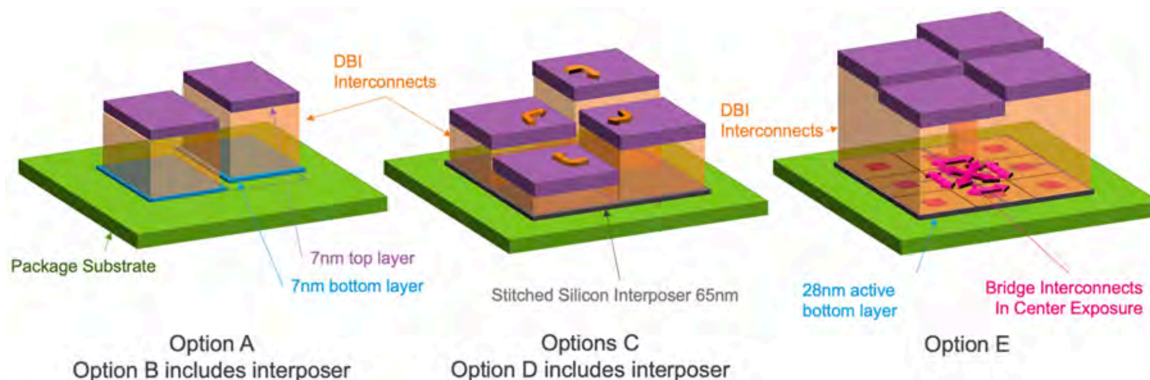


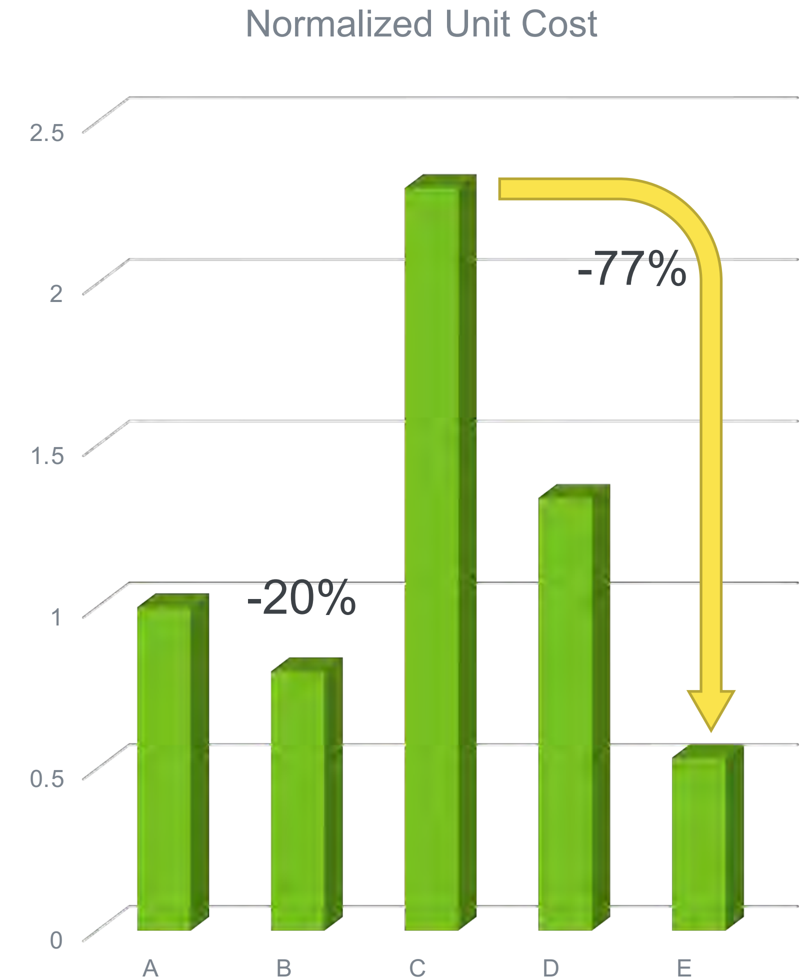


Comparative Unit Cost Analysis

Interface	2.1D + 3D 2 Stacks of 2 Die	2.5D Array of 4 Die
USR	Option A	Option C
HBI	Option B	Option D
Native		Option E

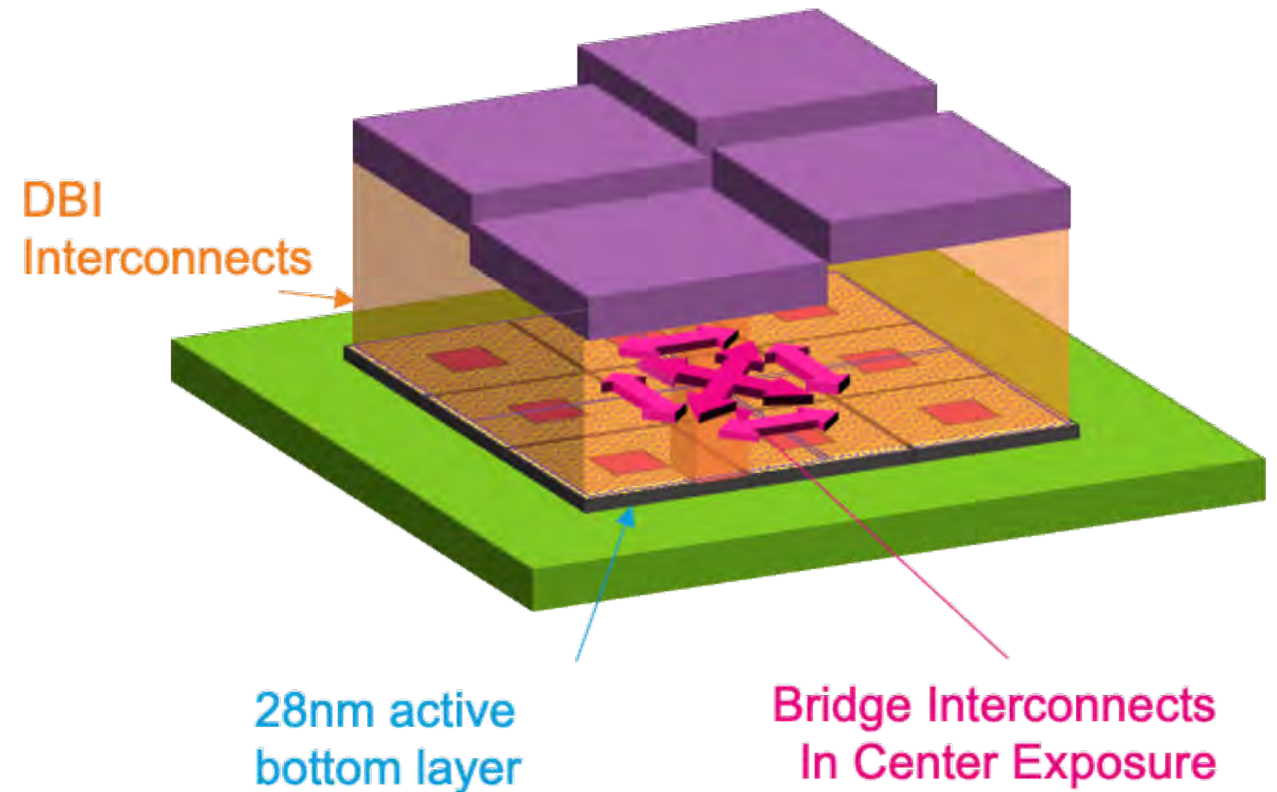
- Reduced total die area improves yield on Option E due to reduced interface area with native interconnects
- HBI is more efficient in space than a USR, but both impact die size





Data Summary

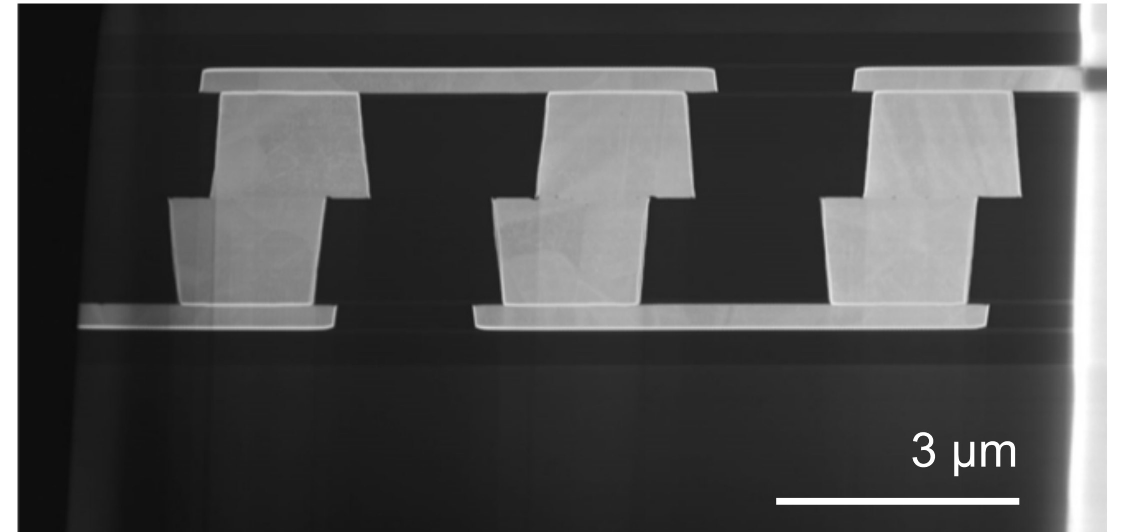
- The most compelling case is option E
  - Lowest interconnect power (-79%)
  - Lowest short route latency (-57%)
  - Lowest unit cost (-77%)
  - Additional mask cost (25%)



XPERI

Summary

- What is the barrier for adoption on this?
- DBI Ultra® die-to-wafer strategies enable new architectures
- Leverage the existing interfaces used within die to span die boundaries.
- 3D allows for a path beyond reticle limits without PPA tradeoffs



STEM from a thin lamella: Z contrast

Acknowledgements: Contributions and PPA analysis performed by Ferran Martorell and Prasad Subramaniam of eSilicon

XPERI®