JLR FINAL SUBMISISON

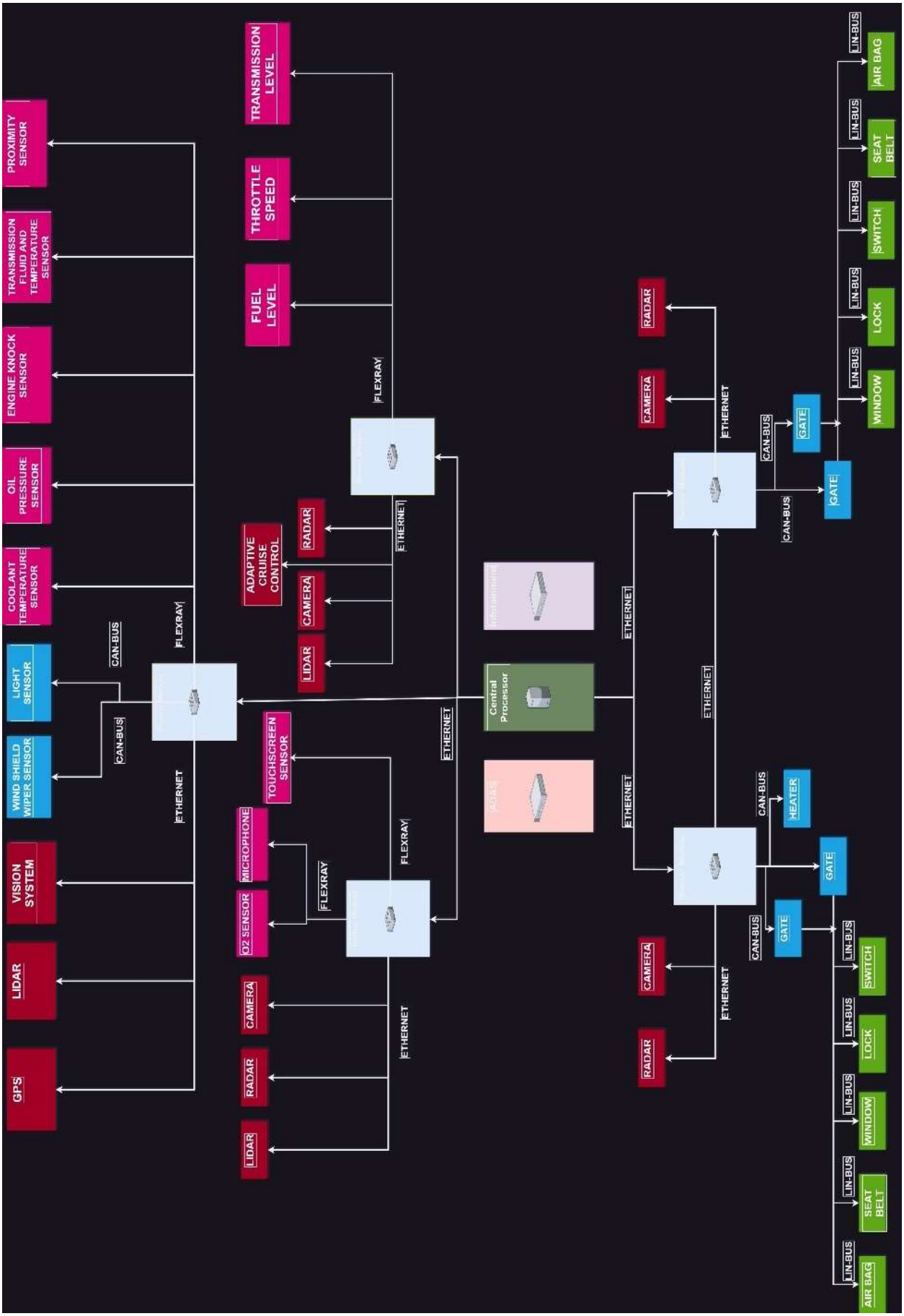
TEAM 46

TABLE OF CONTENTS

LIST OF FIGURES

PREFACE

ZONAL ARCHITECTURE


PROXIMITY SENSOR
TRANSMISSION FLUID AND TEMPERATURE SENSOR
ENGINE KNOCK SENSOR
OIL PRESSURE SENSOR
COOLANT TEMPERATURE SENSOR
LIGHT SENSOR
WIND SHIELD WIPER SENSOR
VISION SYSTEM
LIDAR
GPS
TRANSMISSION LEVEL
THROTTLE SPEED
FUEL LEVEL
ADAPTIVE CRUISE CONTROL
RADAR
CAMERA
LIDAR
TOUCHSCREEN SENSOR
MICROPHONE
O2 SENSOR
CAMERA
RADAR
LIDAR
Central Processor
Infotainment
RADAR
CAMERA
RADAR
CAMERA
RADAR
HEATER
GATE
GATE
GATE
GATE
GATE
AIR BAG
SEAT BELT
WINDOW
LOCK
SWITCH
SEAT BELT
AIR BAG
WINDOW
LOCK
SWITCH
CAN-BUS
FLEXRAY
ETHERNET
LIN-BUS

ZONAL ARCHITECTURE OVERVIEW

Zonal architecture in automotive design involves dividing the vehicle into distinct zones, each serving specific domains or functionalities. This approach enhances modularity, reduces complexity, and enables efficient management of various vehicle systems. Common domains in zonal architecture include:

ADAS (Advanced Driver Assistance Systems):

- Focuses on safety features and driver assistance technologies.
- Involves sensors for collision avoidance, lane departure warning, and adaptive cruise control.

Powertrain:

- Manages the vehicle's propulsion system.
- Includes sensors related to the engine, transmission, and overall powertrain efficiency.

Infotainment:

- Encompasses in-car entertainment, navigation, and connectivity features.
- Involves sensors for touchscreens, cameras, and audio systems.

Climate:

- Controls the vehicle's heating, ventilation, and air conditioning systems.
- Utilizes sensors for temperature, humidity, and air quality.

Body:

- Addresses functions related to the vehicle's structure and safety.
- Includes sensors for airbags, door locks, and chassis integrity.

Zonal Sensor Module Configuration

In the front part of the car, three zonal sensor modules are strategically placed, featuring a star topology for low latency and easy fault tolerance. The configuration includes:

Powertrain Sensors:

- Connected via FlexRay due to matching data rate speeds.
- FlexRay employs TDMA to prevent collisions, ensuring efficient communication.

ADAS and Infotainment Sensors:

- Employ Ethernet for high-speed data transfer.

- Enables seamless connectivity for advanced driver assistance and entertainment systems.

Body and Climate Sensors:

- Utilize CAN-bus and LIN-bus.
- These buses are chosen for their lower data requirements and speeds suitable for body and climate control domains.

At the back of the car, two additional zonal sensor modules are implemented, forming a ring topology for enhanced resilience. This configuration supports reliable communication among sensors and the central processor in the rear part of the vehicle. Overall, the zonal architecture optimizes communication networks based on the specific requirements of each domain, contributing to a more efficient and reliable vehicle system.

LIDAR VS RADAR VS CAMERA VISION

Active and passive sensors serve distinct roles in measuring physical properties and converting them into processable signals. In broad terms, sensors are instruments designed for the measurement of physical attributes, with the capability to convert these measurements into signals suitable for processing, display, or storage.

Active sensors, such as Radar and LiDAR, operate by emitting energy, such as radio waves or laser light, and then measuring the reflected or scattered signals. In contrast, passive sensors, like cameras, identify natural radiation or emissions from the target or the surrounding environment, such as sunlight or artificial light.

Each sensing method has its advantages and drawbacks.

Active sensors, generating their own signals, remain unaffected by external lighting conditions. Radar and LiDAR operate effectively in total darkness and direct sunlight, unlike Cameras. The impact of external lighting on cameras extends beyond night vision, with potential challenges in areas with shadows caused by objects (whether moving or static, like trees or buildings) and even in indoor settings where lighting conditions change (e.g., a door opening introducing more light to the scene).

Weather conditions can significantly impact passive sensors, as their sensing doesn't directly engage with physical phenomena like rain or fog. Instead, they rely on the resulting image, making them more prone to weather-related limitations. Nevertheless, active sensors can also be influenced by adverse conditions, depending on their wavelength.

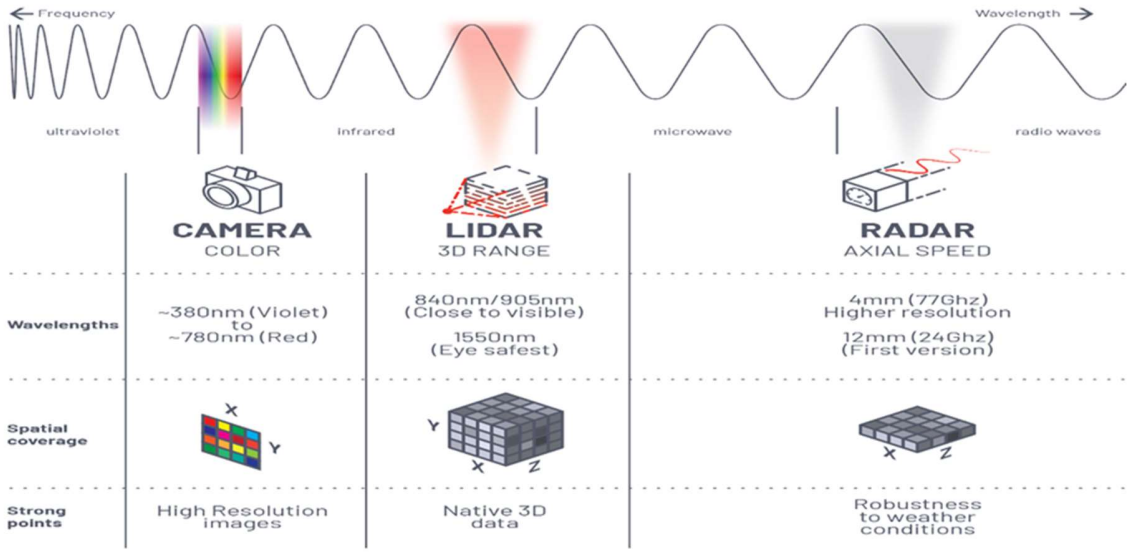


Figure 1 Comparison on the basis of Electromagnetic Spectrum

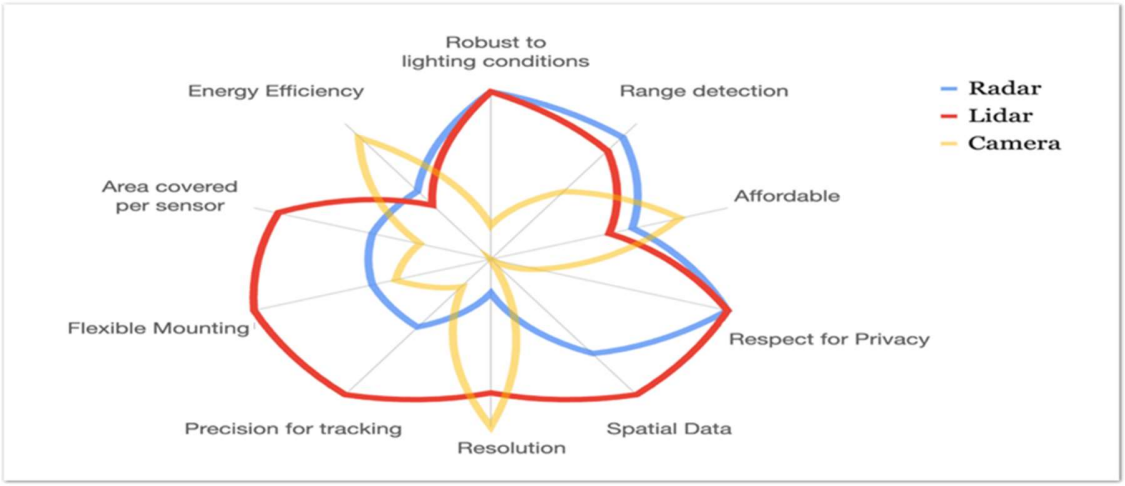


Figure 2 Summary of comparison

From an industry perspective, it is advisable for JLR (Jaguar Land Rover) to consider incorporating all three sensor technologies—LiDAR, RADAR, and Camera vision—in their autonomous vehicle development strategy. This approach aligns with the practices of companies like Waymo, demonstrating the effectiveness of a comprehensive sensor suite in enhancing overall system performance and addressing diverse environmental challenges.

COMPARISONS AMONG DIFFERENT TYPES OF COMMUNICATION PROTOCOLS

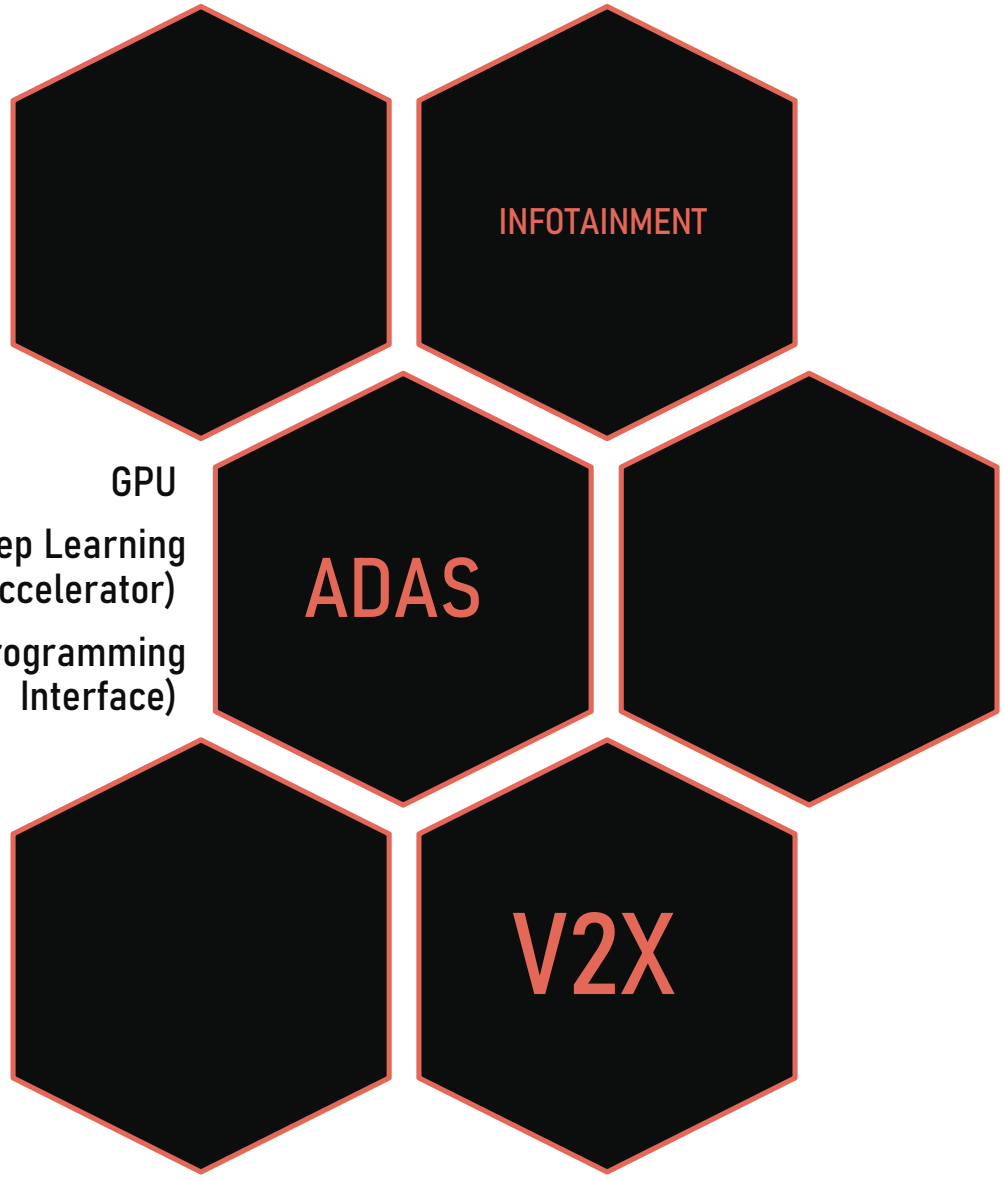
PARAMETERS	CAN	LIN	FLEX-RAY	ETHERNET	PCLE(GEN 4)
Architecture	Multiple nodes (20,32)	Single master & upto 15 slaves	Multiple nodes (upto 64)	Switch based	-
Medium access	CSMA-CR method	Polling method	TDMA method	CSMA	-
Message Transmission	Asynchronous	Synchronous	Synchronous/ Asynchronous	Synchronous/ Asynchronous	-
Data Rate	Max 1Mbps	Max 20Kbps	Max 10Mbps	Max 10Gbps	Max 16Gtps
Bit Coding	NRZ and bit stuffing	NRZ	NRZ	NRZ	128b/130b
Physical Layer	Electrical dual wire	Single Electrical wire	Dual wire Optical or Electrical	Twisted pair cables	-
Range	40 m	1-5km	10 m	Max 25m	-

While developing and physically realizing our Micro-Architecture Diagrams, we must keep throughput and robust communication as our paramount priority.
Throughput refers to the rate of useful work done, or information processed over a period. It is a measure of efficiency and productivity of our given design.

In order to maximize throughput, the following principles will be abided:

- **Parallel Processing**: Harness the power of parallelism to concurrently execute multiple tasks, whether through task parallelism, data parallelism, or instruction-level parallelism.
- **Pipelining Structure**: Employ a pipeline architecture that divides task processing into stages, facilitating continuous execution and boosting overall throughput.
- **Concurrency Management**: Design systems capable of handling multiple tasks concurrently, allowing independent processes to progress simultaneously without interference.
- **Scalable Design**: Create architectures that seamlessly scale with increasing workloads, accommodating additional resources like processors or servers without compromising performance.
- **Load Distribution**: Ensure even distribution of tasks across available resources to prevent
- **Caching Mechanisms**: Implement efficient caching strategies to store frequently accessed data, reducing the need for retrieval from slower memory sources and enhancing system responsiveness.
- **Memory Hierarchy Optimization**: Design a memory hierarchy that effectively uses various memory levels (registers, cache, RAM) to minimize data retrieval times and improve overall system speed.
- **Asynchronous Communication**: Use asynchronous processing and communication to decouple tasks, minimizing idle time and enhancing system efficiency.
- **Predictive Analysis**: Integrate predictive analysis and prefetching mechanisms to expect future data and instruction needs, reducing latency and boosting throughput.
- **Bandwidth Management**: Optimize data transfer bandwidth by minimizing data movement, using efficient communication protocols, and optimizing network architecture.
- **Fault-Tolerant Design**: Incorporate fault-tolerant features to mitigate the impact of system failures or errors on overall throughput. This may involve redundancy, error-checking mechanisms, and graceful degradation.
- **Algorithmic Efficiency**: Select algorithms tailored to the parallel and distributed nature of the system, ensuring effective processing of tasks.
- **Resource Monitoring**: Continuously monitor and analyze resource utilization to identify bottlenecks, underutilized resources, and areas for improvement.
- **Power-Efficient Architectures**: Consider power consumption and design architectures with energy efficiency in mind, especially in scenarios where power efficiency is critical.

WHAT PARTS WE HAVE COVERED IN THE MICROARCHITECTURE DIAGRAMS



GPU

DLA (Deep Learning Accelerator)

VPI (Visual Programming Interface)

ADAS DOMAIN CONTROLLER

For our ADAS/Infotainment Architecture, we analyzed the specifications and architectures of the Key players in Market such as AMD, Snapdragon and Nvidia.

Nvidia has announced a new ADAS Domain Unit called the Nvidia Thor which will be released soon; below however we have compiled the specifications of Nvidia Orin, that is currently the latest generation of Nvidia's AGX units.

GPU (Graphics Processing Unit)

1. Ampere Architecture:
   1. GPCs (Graphic Processing Clusters)
   2. TPCs (Texture Processing Clusters)
   3. SMs (Streaming Multiprocessors)
   4. CUDA Cores and Tensor Cores
   5. 3rd Generation Tensor Cores

CPU (Central Processing Unit)

1. Arm Cortex-A78AE
2. 12-core Configuration
3. L1, L2, and L3 Caches

AI Accelerators

1. Deep Learning Accelerator (DLA):
   1. NVDLA 2.0
   2. Structured Sparsity Support
   3. Depth Wise Convolution
2. Vision Processing:
   1. PVA v2 (Programmable Vision Accelerator)
   2. VIC 2D Engine (Video Imaging Compositor)
   3. VPI Library for Vision Programming Interface

Memory:

1. 256-bit LPDDR5 (32GB or 64GB)
2. 64GB eMMC 5.1 Storage

I/O Interfaces:

1. PCIe Gen4 (22 Lanes)
2. USB 3.2 (Multiple Ports)
3. Gigabit Ethernet
4. 10 Gigabit Ethernet
5. MIPI CSI-2 (16 Lanes)
6. HDMI/DP (Display Port)
7. UART, SPI, I2S, I2C, CAN
8. GPIOs

Video Codecs:

1. NVENC (Video Encoder)
2. NVDEC (Video Decoder)
3. NVJPEG (JPEG Processor)

Power Management:

1. PMIC (Power Management Integrated Circuit)
2. Voltage Regulators
3. Power Profiles (15W to 60W)

Software Support:

1. JetPack SDK
2. TensorRT, cuDNN
3. DeepStream, Isaac, Riva SDKs
4. TAO Toolkit and Pre-Trained Models

Figure 3 Specification of Nvidia Thor

Our Architecture proposed a hybrid Processing Network

- Application Processing – Less Power Sensitive Tasks
- A dedicated Processing Unit – Power Sensitive Tasks

This also employs various features for security and Cryptography that you will shortly see in the deeper levels of our Micro-Architecture.

LEVEL 1

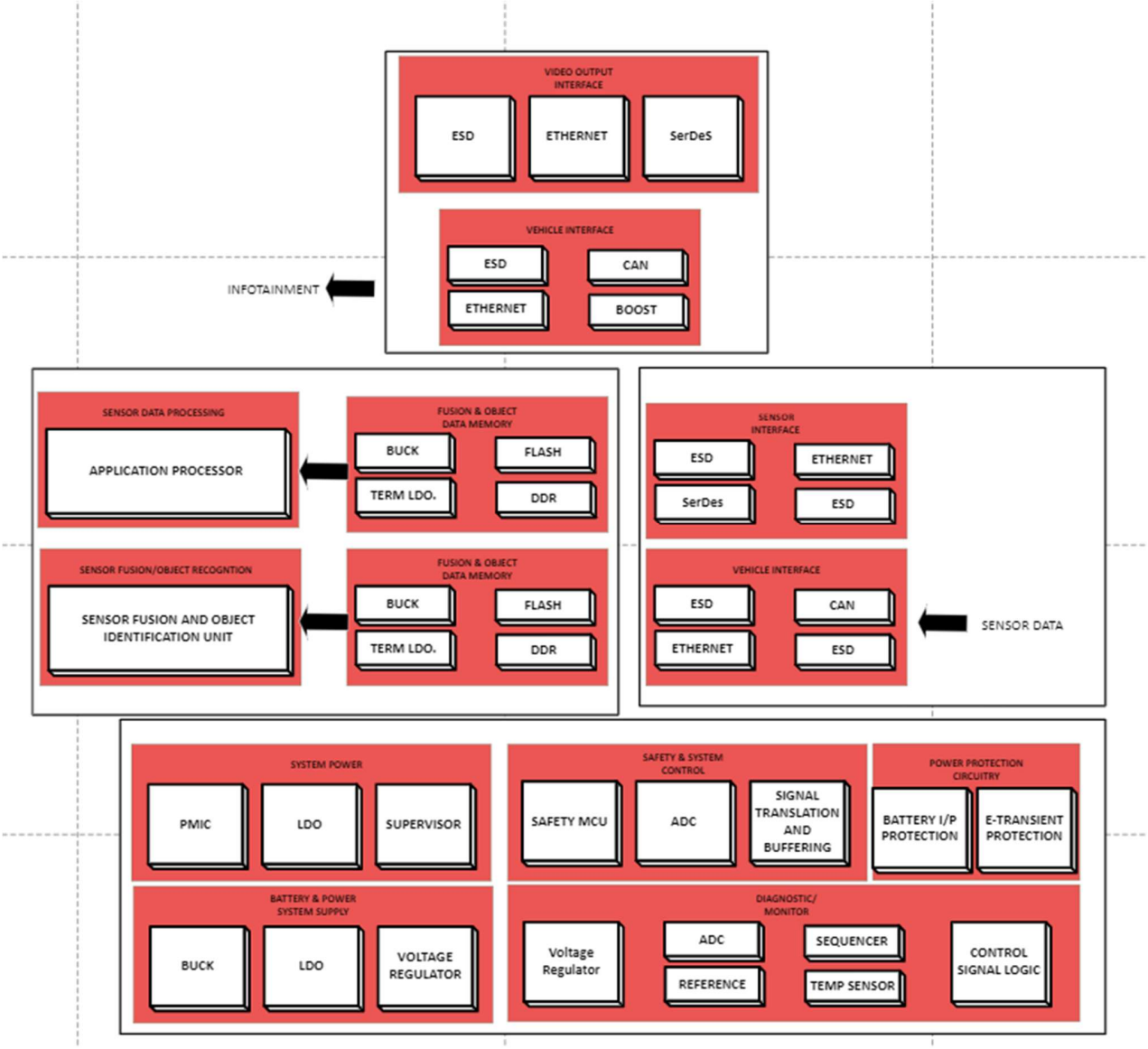


Figure 4 ADAS Domain Controller Architecture Diagram

LEVEL 2

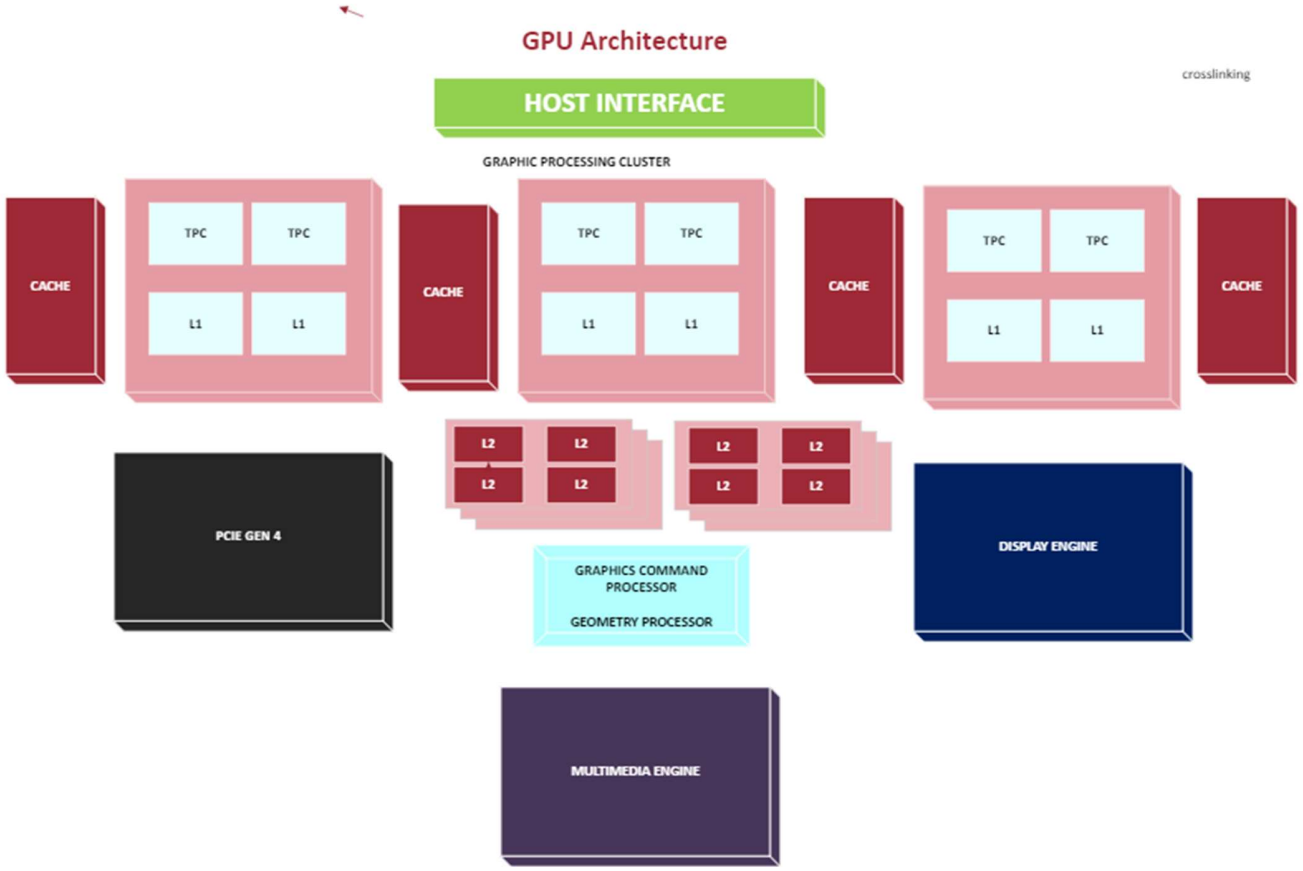


Figure 5 GPU Architecture Diagram

For our GPU Architecture we have used the **Cross-Linking Technique** by placing Cache's in between of Graphical Processor Clusters, this is a technique used by AMD to enable faster access to Cache and make it monolithic from a software developer's perspective.
Here we demonstrate both a monolithic and a chiplet based GPU Design in multiple levels of depth. Chiplet based GPU (MCM-GPU) can unlock increased levels of performance despite the limitations in transistor scaling. To maximize throughput in this architecture we have followed the following principles:

1. Minimize Inter GPU-Module Traffic
2. We will utilize local hardware Cache to capture Local traffic in the GPU Module
3. Utilizing Cooperative Thread Array (CTA)
4. First Touch page allocation Policy to minimize Inter GPM (GPU modules) Traffic

MCM (Multi Chip Module) GPU chiplets excel over multi-GPU setups, ensuring superior bandwidth, energy efficiency, scalability, and reduced latency for optimal performance.
As we can observe with the measures given below, Nvidia is also moving towards the same trends in iterations of their GPU Architecture.

Dr. Arkprava Basu

Associate Professor, IISc Bangalore

Data travelling from one die to another takes about 32ns. We must keep in mind for the architecture that we minimize the access to L2 Cache that is remote and design such that we map virtual addresses and limit number of remote cache lookups. Such a measure has been able to speed up data fetching in MCM designs by 52% on average

	Fermi	Kepler	Maxwell	Pascal
SMs	16	15	24	56
BW (GB/s)	177	288	288	720
L2 (KB)	768	1536	3072	4096
Transistors (B)	3.0	7.1	8.0	15.3
Tech. node (nm)	40	28	28	16
Chip size (mm2)	529	551	601	610

Figure 6 Statistics for Nvidia's different generations GPU

The Graphical Processing clusters in the GPU's are composed of Streaming Multiprocessors. We will come across them in many architectures let's discuss what they are

"A GPU is composed of SMs, and each SM contains several SPs. Currently there are 8 SPs per SM and between 1 and 30 SMs per GPU, but really the actual number is not a major concern until you're getting advanced."

- Tom, NVIDIA Developer

LEVEL 3

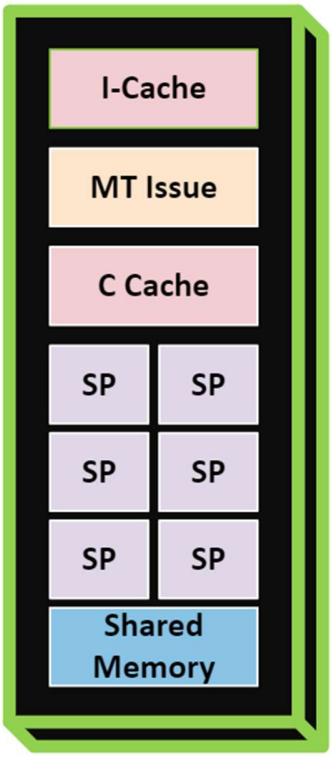


Figure 7 Inner Block Diagram of Streaming Multiprocessor inside a GPU

Here are some of the features in our Micro-Architecture Design. We have explained the reasons and functionalities of each module down to the Streaming Multiprocessor

Feature	Description
Context Switching	Efficient, no cost, every cycle
Instruction Set Architecture (ISA)	Scalar Register-based, versatile, efficient
Multithreading	Up to 1024 concurrent threads, Multithreaded Instruction Unit
Precision Computing	8 SP (single-precision) with IEEE 754 32-bit float, 32-bit/64-bit integer
Thread Management	Hardware thread scheduling, in-order issue for performance
Specialized Function Units (SFU)	2 SFUs for complex math (sin, cos, log, exp)
Double Precision Dominance	Double Precision Unit with IEEE 754 64-bit float and fused multiply-add
Shared Memory Hub	16KB Shared Memory for efficient thread communication

Figure 8 Feature-Description of our proposed Microarchitecture Diagram

V2X

It is widely acknowledged that the Internet of Vehicles (IoV) is a useful tool for improving the dependability, security, and environmental impact of road traffic. There are now two main categories of IoV technologies in the world:

- Cellular vehicle-to-everything (C-V2X)
- Dedicated short-range communication (DSRC), which is based on IEEE 802.11p.

Long Term Evolution (LTE)-V2X and 5G New Radio (NR)-V2X are included in C-V2X. Developed nations and regions—including the US, Japan, and Europe—place a high priority on the development of the Internet of Vehicles. They actively participate in technical research and validation projects that are pertinent.

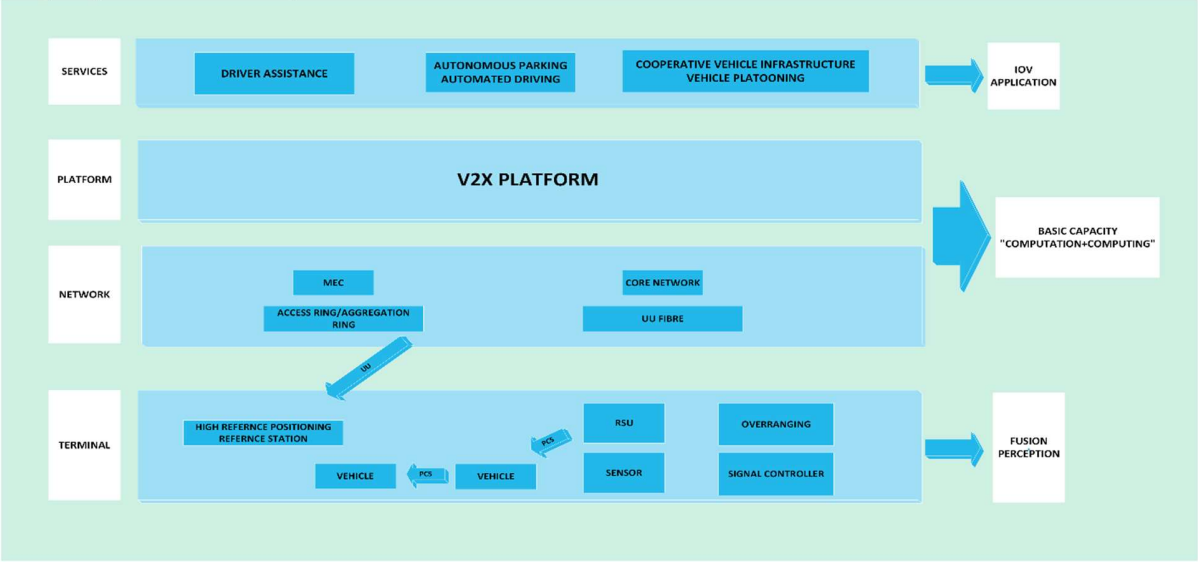


Figure 9 The IoV's overall V2X architecture

The Internet of Vehicles (IoV) is largely dependent on the V2X platform, which provides the necessary data capabilities to support upper-layer applications and V2X services..
These capabilities include

- High concurrent access
- Fused analysis
- High-precision positioning
- Network capability openness
- Edge computing
- Ensuring business continuity

The platform is made to fulfill the functional needs of autonomous driving and driving assistance.

V2X services have unique characteristics that include high concurrency, instantaneous response, swift mobility, heterogeneous data, and infrastructure sharing

High Concurrency:

- Five main categories of information are used in data estimate in accordance with national and international standards (BSM, SPAT, MAP, RSI, and RSM)..
- Concurrent access and computation of millions of data are handled at the district and county levels, tens of millions at the city level, and hundreds of millions at the national level.

High Real-time:

- Following industry standards, the delay for driver assistance does not exceed 100 ms, and the delay for automated driving should not exceed 20 ms.

High-speed Mobility:

- In order to provide continued operations during a vehicle's movement between distinct V2X edge service nodes, V2X services must allow smooth business processing between edge service nodes while supplying edge access and compute services.

Data Heterogeneity:

- IoV service data exhibits a variety of heterogeneities, including standard datasets (BSM, MAP, SPAT, RSM, and RSI) as well as non-standard data (millimeter-wave radar, camera acquisition, LIDAR, vehicle navigation, etc.).

Infrastructure Sharing:

- The deployment of IoV applications requires a fundamental environment since IoV infrastructure is diverse and expensive to build. This setting optimizes the use of vital resources by enabling real-time data access, transmission, and shared processing capacity.

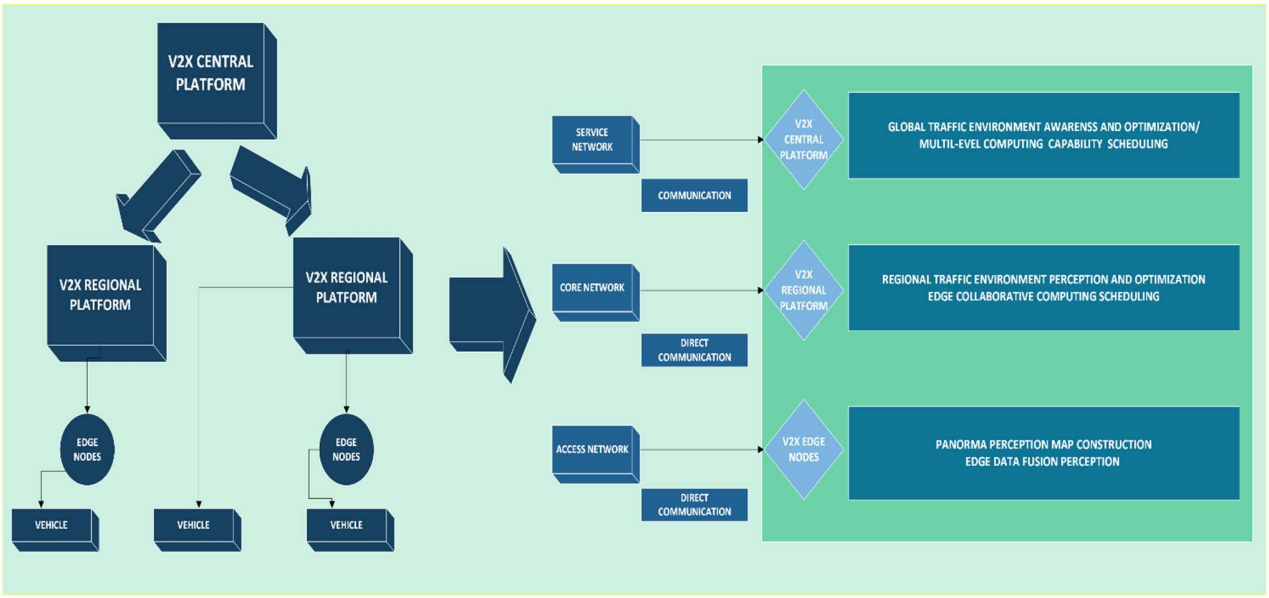


Figure 10 V2X platform deployment across a network

V2X Central Platform:

1. Offers services for the entire network.
2. Offers services for global management.
3. Contains control of network-wide service operations.
4. Oversees the optimization and knowledge of the global traffic environment.
5. Has scheduling for multi-level computing capability.
6. Enables dynamic multi-level deployment of applications.
7. . Oversees data and services that span regions.

V2X Regional Platforms:

1. Primarily assists local and provincial businesses.
2. Has to do with managing regional service operations.
3. Oversees the optimization and perception of the local traffic environment.
4. Contains opening, application hosting, and regional data analysis.
5. Enables scheduling of edge collaborative computing.
6. Oversees the V2X edge node.

V2X Edge Node:

1. Offers high real-time and high bandwidth V2X services primarily in the edge range.
2. Offers features like perception of edge data fusion at the edge range.
3. Makes it easier to generate dynamic panoramic perception maps.
4. Applications for servers

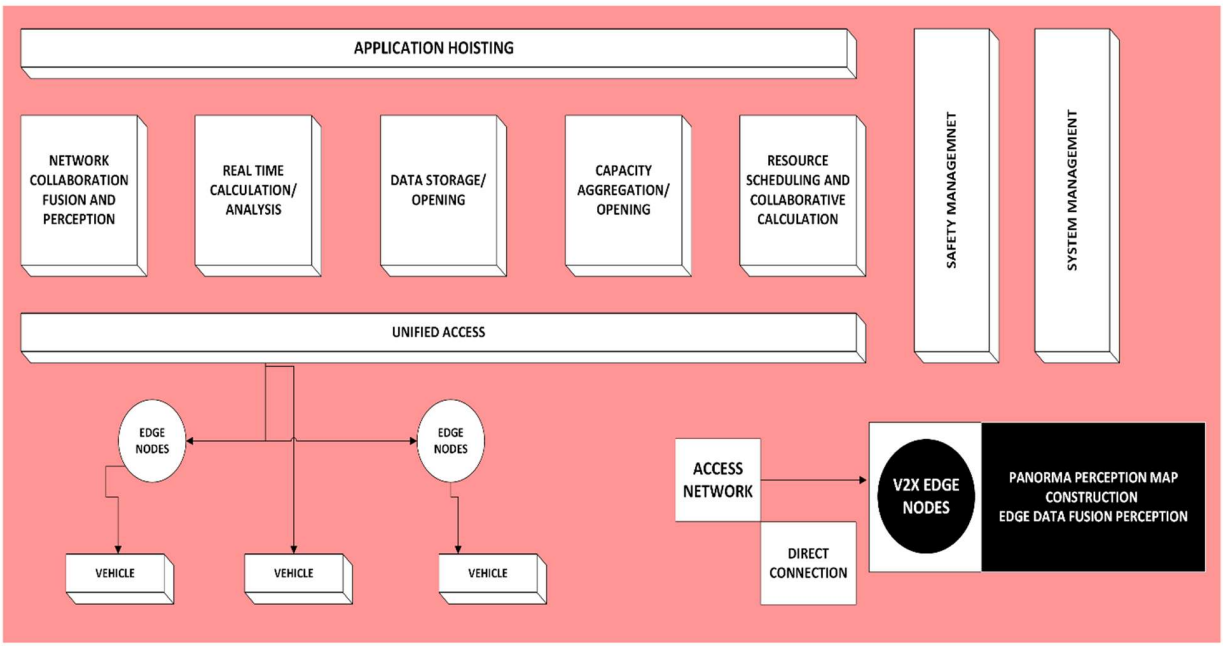
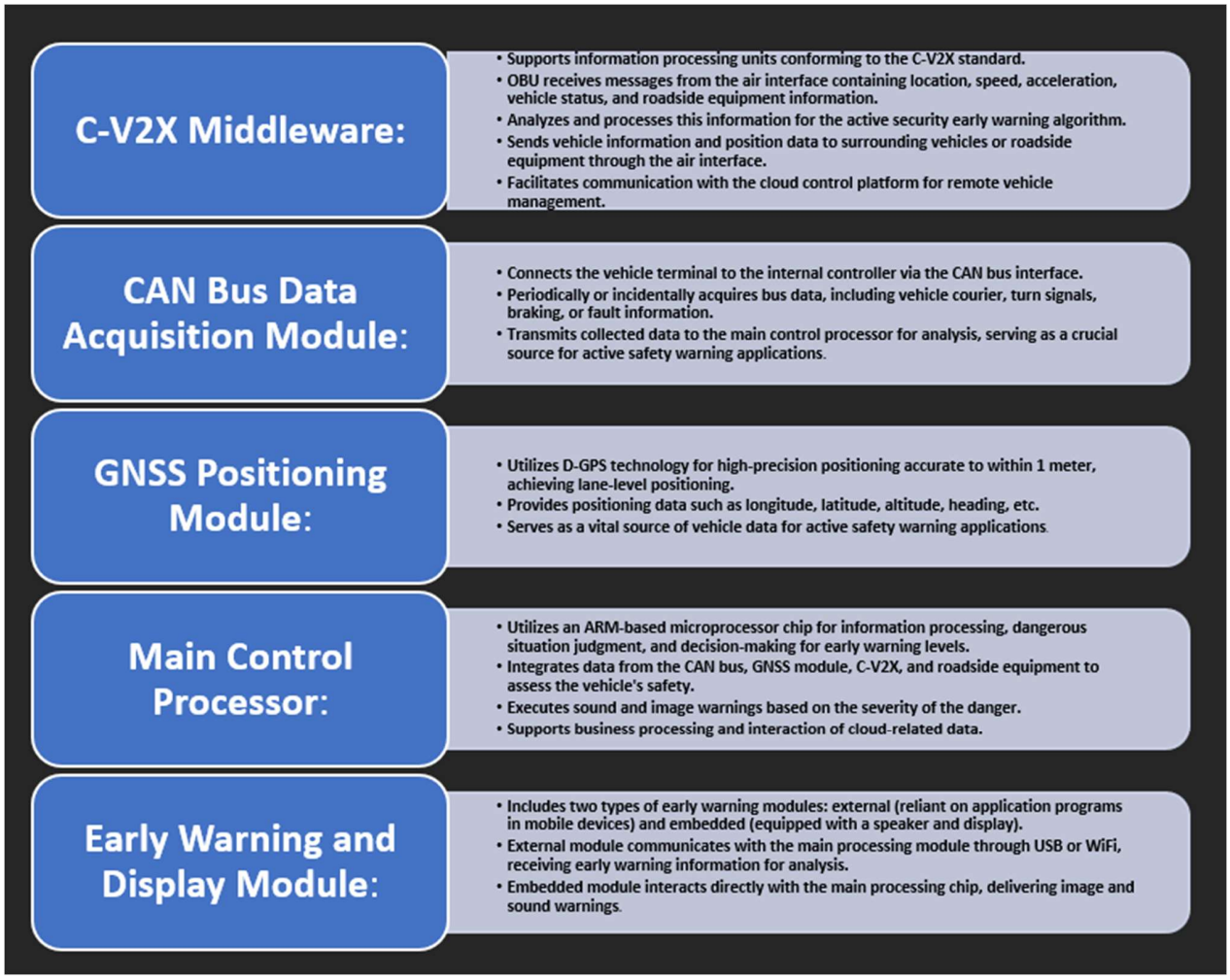


Figure 11 V2X platform functional architecture diagram.

The following is a description of the IoV intelligent IBU hardware's core modules:



The intelligent on-board unit (OBU) hardware design framework based on 5G/LTE-V2X technology is depicted in this picture.

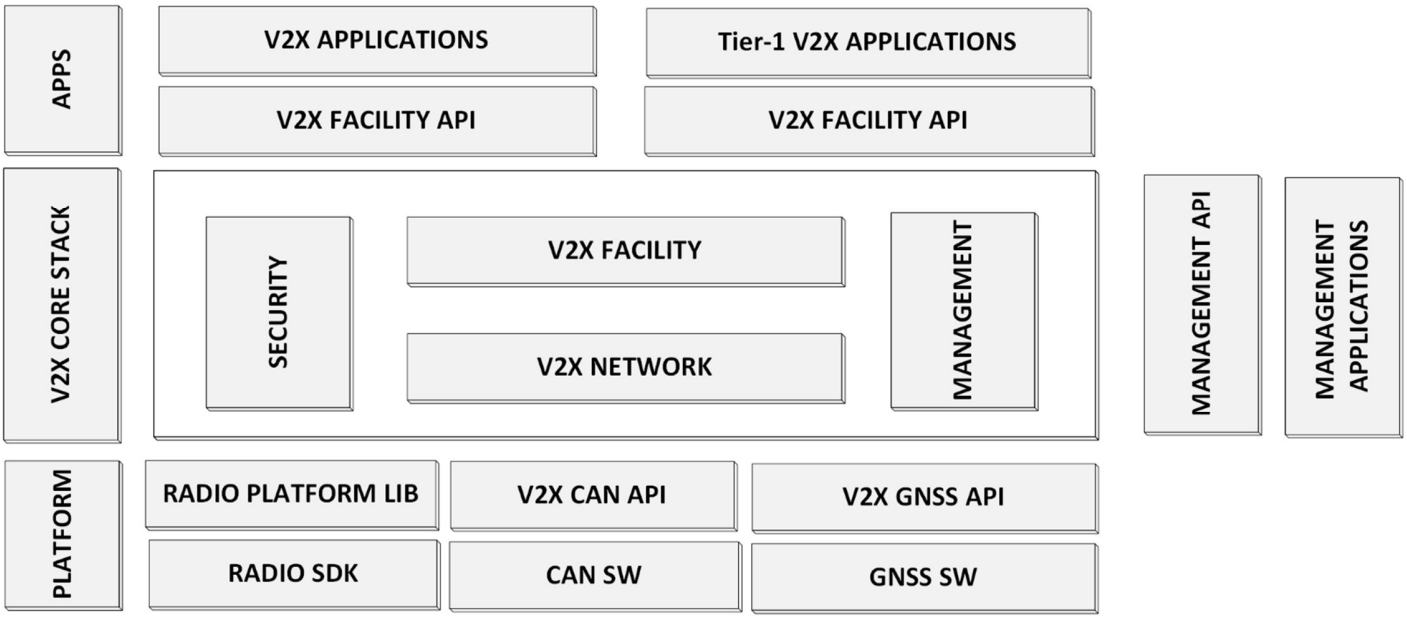


Figure 12 Block schematic shows the hardware architecture for the 5G-based vehicle terminal OBU.

RSU ARCHITECTURE UTILIZING C-V2X AND 5G

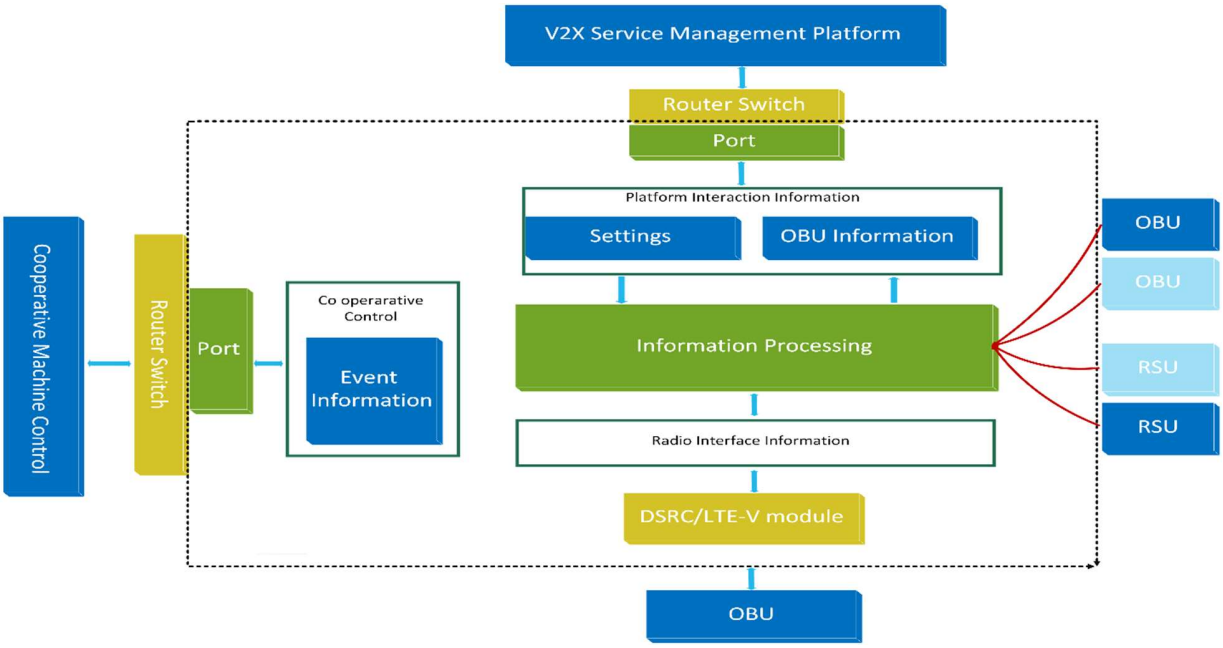
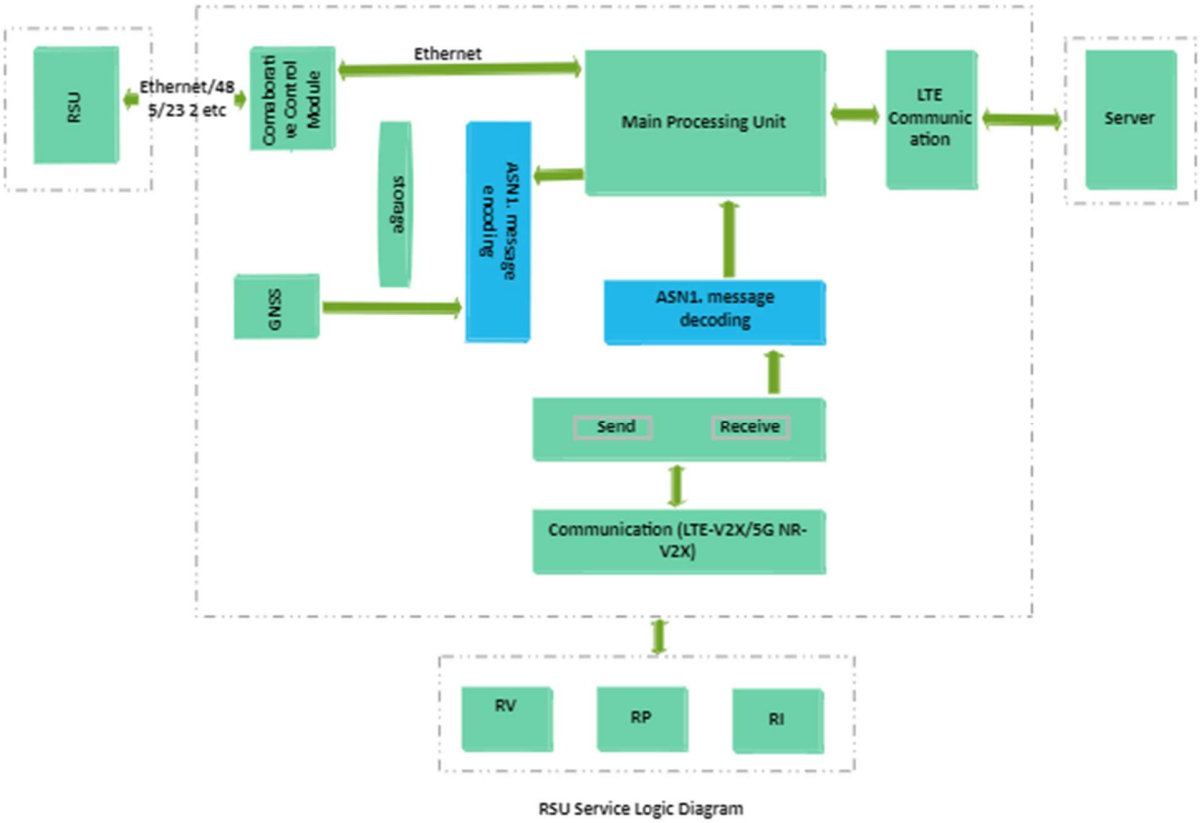


Figure 13 Diagram of the RSU service logic

The diagram shows that the RSU can link to the roadside Mobile Edge Computing Unit (MEC) in addition to acting as a V2X communication channel. In turn, this MEC has access to a range of roadside traffic sensing and monitoring devices, such as variable message signs, smart cameras, laser, microwave, and millimeter-wave radars; in addition, environmental perception tools, road surface sensors, and signal light systems are also available. Together, these gadgets keep an eye on motor vehicles, non-motor vehicles, pedestrians, and other road users in order to identify and track traffic-related incidents such as accidents, bottlenecks, and stray animals. They also help with functions like assessing the condition of the pavement and perceiving the rain and fog. The system accomplishes real-time processing and dynamic collaboration of local traffic data by fusion processing of the collected data, hence eliminating delays.



RSU Service Logic Diagram

Figure 14 Diagram of RSU Service Logic

The following is included in the RSU Service Logic Diagram:

1. **LTE-V2X/5GNR-V2X Communication Module**: This module allows the RSU to communicate with surrounding vehicles, pedestrians, and other RSUs by allowing messages to be exchanged.

2. **N.1 Message Protocol Data**:
   The RSU modifies traffic facilities in accordance with applicable control requirements after analyzing received ASN.1 national standard message protocol data.

3. **NSS Module**:
   The GNSS module sets positional parameters such mode, output format, frequency, and baud rate. It is used to locate and time roadside terminals. It uses dual-mode BSD and GPS for positioning, and the output format is NMEA compliant.

4. **Collaborative Control Module:** This module is used by the RSU to collect traffic facility data and encodes it using the national message dictionary standard ASN.1. The sending module then arranges, transmits, and broadcasts the data in real time. This comprises RSI messages for roadside electronic signage, MAP messages with data from road sections, and RSI messages with data from traffic lights, pedestrians, and machines for participants in traffic and incidents on the road.

Road traffic facility signal collecting is made possible by the dedicated Ethernet port connecting the RSU to the collaborative control module. This includes variable message signs, traffic light signals, roadside signs, and traffic monitoring devices. It also supports RSU applications including road danger warnings, speed guiding, signage, congestion alerts, and red light alerts.

INFOTAINMENT SYSTEM

Why Introduce Augmented Reality?

In 2019 the Minister of Road Transport & Highways (MoRTH), Nitin Gadkari said that the government would not push for fully autonomous transportation as it would leave millions unemployed. The following issues are observed in introducing autonomous driving on indian roads:

- Secure autonomous systems are dependent on road infrastructure. The lack of signages and road markings with the addition of potholes make an accurate fully autonomous system a distant reality.
- The biggest enemy to an autonomous car is non-autonomous vehicles. The unregulated nature of Indian streets with reckless driving and ignorance of rules, puts autonomous vehicles at a disadvantage.
- Indian laws do not allow for testing of prototypes on Indian streets. Car companies are forced to run prototypes in controlled environments.
- Suburban and rural roads are also prone to animal obstructions. This is seen at a much higher rate in India.
- Motor Vehicles Act, 1988, Consumer Protection Act, 2019, Information Technology Act, 2000 - are in particular posing legal challenges in adoption of autonomous vehicles.

The question is if these hindrances are holding back the adoption of fully automated driving, how do we move forward in the transportation industry. The approach we propose is to instead move towards improving the experience and performance of the driver. While maintaining a level 2 autonomy with the addition of Augmented Reality we can offer driver assistance and user services at a much better rate. Augmented reality may enhance the drivers performance and improve transportation both at a consumer and business level.

POTENTIAL OF AUGMENTED REALITY IN CARS

Intelligent Navigation

Figure 15 Theoretical animation of AR implementation by Samsung

Figure 16 AR Implementation in HUD

The combination of Augmented reality and Heads Up Displays (HUD) allows car manufacturers to provide a truly driver focused assistive feature. Instead of multiple head movements and additional devices to help navigate new terrains. This feature simplifies the driving experience, all the information required by the driver is available right in front. It offers a safer driving experience coupled with reduced driving strain.

- Directions and terrain information is available on screen.
- Travel speed and any traffic regulations are provided on screen.
- Recommended driving strategies, like lane changes, braking zones, where to park, recommended speeds improve the safety and reduce driving strain.

Information

- AR gives us the ability to display all relevant information in an intuitive way. Any information required by the user can be displayed as shown.
- Locations of restaurants, hotels, stores and other businesses can be advertised for potential customers. This can be a source of business for JLR.
- The feature can also be used to improve the experience inside the car. Onboarding process becomes easier, and a wide range of additional features can be accommodated.

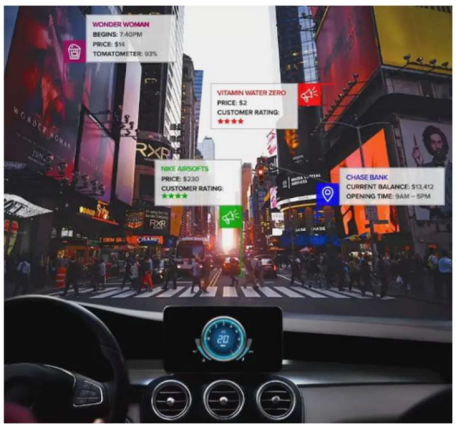


Figure 17 Displaying inforamtion on HUD

- Historical sights, landmarks and other tourist centered information can be displayed allowing users to navigate new areas and provide an ideal travel experience.

Safety

Figure 19 Low visibility conditions

Figure 18 AR based Pedestrian detection

Safety is critical in developing driving systems. Augmented reality allows us to use radar and lidar feeds to display road obstructions, pedestrians, animals, lane markings and signs in low visibility conditions. This improves the performance of older and less experienced drivers. This is a feature that would appeal to modest drivers and plays a critical role in road safety.

Types of Implementations

The two major players in display technology are Heads up display (HUD) and Heads down display (HDD). Regular studies have been done comparing the user experience regarding safety and functionality. The studies listed discuss the effects on users. Key takeaways :

- HUDs improve the driving performance of the average consumers.
- Drivers prefer to use a HUD rather than a HDD.
- Placement of the display plays a key role in effectiveness, HUDs require less head and eye movements. With increase in distractions in the cabin, It would make more sense to move towards a HUD approach.
- Some studies show negligible improvement with HUD in place, but in these studies seasoned drivers like truck drivers instead of the average consumer.

Features in Infotainment system

SAFETEY & FUNCTIONALITY	ENTERTAINMENT
Vehicle diagnostics	Display system
AR based driving assistance	Audio system
Parking assistance	Phone connectivity
Communication	Cabin Climate Control

Trends in Automotive Infotainment Systems

As technology progresses, IVI systems are integrating larger and more prominent screens with increased resolutions, aiming to offer visually captivating and user-friendly interfaces. The adoption of larger, higher-resolution displays opens up possibilities for split-screen features, enabling the simultaneous display of multiple applications or information. For instance, drivers can now observe maps and control their music concurrently or have navigation instructions presented alongside climate control settings.

New age infotainment systems are powered by powerful automotive processors designed for advanced IVI systems. These automotive processors are capable of displaying content on multiple displays (e.g. Head-up Display or Windshield, Connected smartphones, Head Unit, and more) and deliver an enhanced in-vehicle experience to drivers and passengers.

The advanced infotainment system is designed to seamlessly integrate with multiple screens and displays, offering a versatile and immersive user experience. With a comprehensive array of features, the system is expected to support up to seven screens of varying resolutions and sizes, strategically placed for optimal usability: three on the front console, two on the rear side, a cutting-edge head-up display featuring augmented reality, and a digital rearview mirror.

The system leverages multiple GPUs (Graphics Processing Units) and DSP (Digital Signal Processing) processors to ensure a smooth and efficient operation. This combination of hardware

enhances the overall performance and prioritizes safety features, guaranteeing a secure and reliable driving experience. The intricate architecture of the system involves the integration of multiple CPUs, GPUs, and DSP processing units.

Infotainment system design

Key design choices:

- Two C7x floating-point, vector DSPs, up to 1.0 GHz, 160 GFLOPS, 512 GOPS.
- Deep-learning matrix multiply accelerator (MMA), up to 8 TOPS at 1.0 GHz.
- Vision Processing Accelerators (VPAC) with Image Signal Processor (ISP) and multiple vision assist accelerators.
- Depth and Motion Processing Accelerators (DMPAC).
- Dual 64-bit Arm Cortex-A72 microprocessor subsystem at up to 2 GHz.
- Up to six Arm Cortex-R5F MCUs at up to 1.0 GHz.
- GPU IMG BXS-4-64, 256kB Cache, up to 800 MHz, 50 GFLOPS, 4 GTexels/s.

Memory Subsystem:

- Up to 4MB of on-chip L3 RAM with ECC and coherency.
- Two External Memory Interface (EMIF) modules with ECC, supporting LPDDR4 memory types.
- General-Purpose Memory Controller (GPMC).
- On-chip SRAM in MAIN domain, protected by ECC.

Device Security:

- Secure boot with secure runtime support.
- Customer programmable root key, up to RSA-4K or ECC-512.
- Embedded hardware security module.
- Crypto hardware accelerators – PKA with ECC, AES, SHA, RNG, DES, and 3DES.

High-Speed Serial Interfaces:

- PCI-Express Gen4 controllers (up to four lanes per controller).
- USB 3.0 dual-role device (DRD) subsystem with enhanced SuperSpeed Gen1 Port.
- CSI2.0 4L RX plus CSI2.0 4L TX.

Display Subsystem:

- DSI 4L TX, eDP 4L, DPI.

Audio Interfaces:

- Five Multichannel Audio Serial Port (MCASP) modules.

Ethernet:

- Two RMII/RGMII interfaces.

Figure 20 Complete Infotainment Architecture

Component	SoC node size*	Bought/In house	Suggested	Chiplet node size
Main computer	11nm	Bought	ARM CPU by TI	11nm
GPU	11nm	In House	(By JLR)	11nm
DSP	11nm	Bought	TI DSP	28nm
Display Subsystem	11nm	In House	(By JLR)	28nm
Audio Subsystem	11nm	In House	(By JLR)	28nm
Camera Subsystem	11nm	In House	(By JLR)	28nm

* (Qualcomm SA6155P, n.d.)

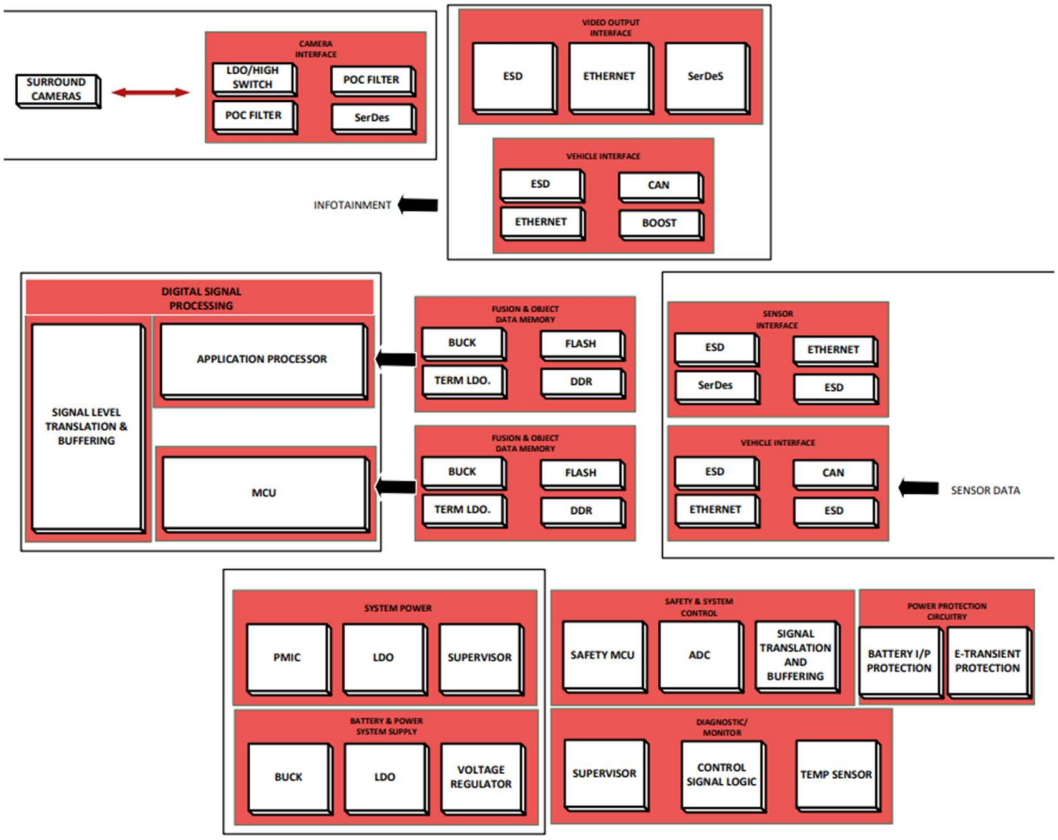


Figure 21 Model Architecture for Infotainment

Processor Cores	**C7x DSPs**	**2 Floating-point, Vector**
	Clock Speed	Up to 1.0 GHz
	Performance	160 GFLOPS, 512 GOPS
	MMA Accelerator	Up to 8 TOPS at 1.0 GHz
	VPAC with ISP	Yes
	DMPAC	Yes
	Arm Cortex-A72	Dual 64-bit at up to 2 GHz
	Arm Cortex-R5F MCUs	Up to 6 at up to 1.0 GHz
	GPU	IMG BXS-4-64
	GPU Cache	256kB
	GPU Clock Speed	Up to 800 MHz
	GPU Performance	50 GFLOPS, 4 GTexels/s
Memory Subsystem	On-chip L3 RAM	Up to 4MB with ECC and coherency
	External Memory Interfaces (EMIF)	2 with ECC, LPDDR4
	General-Purpose Memory Controller (GPMC)	Yes
	On-chip SRAM (MAIN domain)	Yes, with ECC
Functional Safety	Compliant (select part numbers)	
	Developed for ISO 26262 applications	Yes
	Hardware Integrity (MCU Domain)	Up to ASIL-D/SIL-3
	Hardware Integrity (Main Domain)	Up to ASIL-B/SIL-2
	Safety-related certification	Planned for ISO 26262
Device Security	Secure boot	Yes
	Customer programmable root key	Up to RSA-4K or ECC-512
	Embedded hardware security module	Yes
	Crypto hardware accelerators	PKA with ECC, AES, SHA, RNG, DES, 3DES
High-Speed Serial Interfaces	PCI-Express Gen4 controllers	Up to 4 lanes per controller
	USB 3.0 DRD	Yes
	CSI2.0	4L RX + 4L TX
Automotive Interfaces	MCAN modules	20 with full CAN-FD support
Display Subsystem	DSI 4L TX, eDP 4L, DPI	Yes
Audio Interfaces	MCASP modules	5 Multichannel
Ethernet	RMII/RGMII interfaces	2
Flash Memory Interfaces	eMMC 5.1	Yes
	SD 3.0/SDIO 3.0	Yes

Display Subsystem

With the increase in displays in the car cockpit, significant computational resources are utilized in running these displays. Since the use of multiple displays is a limited need for certain use cases, it would make a lot of sense to manufacture the DSS in house, i.e. it is our recommendation that JLR manufacture this chiplet. Inspiration was taken from Texas Instruments' application in the Jacinto 7 Infotainment system. The DSS performs multi-layer composition for the display output and supports a set of industry standard display interfaces to drive a wide range of display panel resolutions.

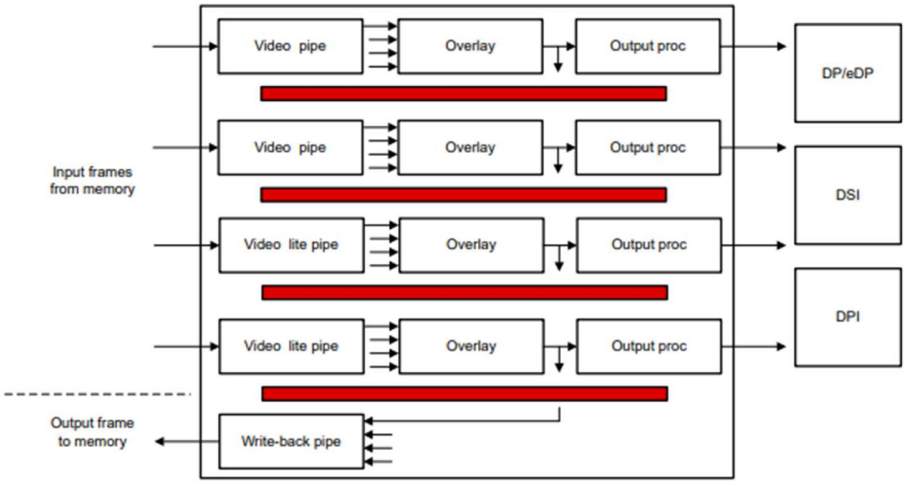


Figure 22 Display Subsystem Architecture

The various display interfaces can be made compatible by modifying the display subsystem. This gives us a solution to the growing display requirements in the autonomous car. Alleviating the processing load of the main processor and using design specific (based on display type and number) reduces the overall resources used.

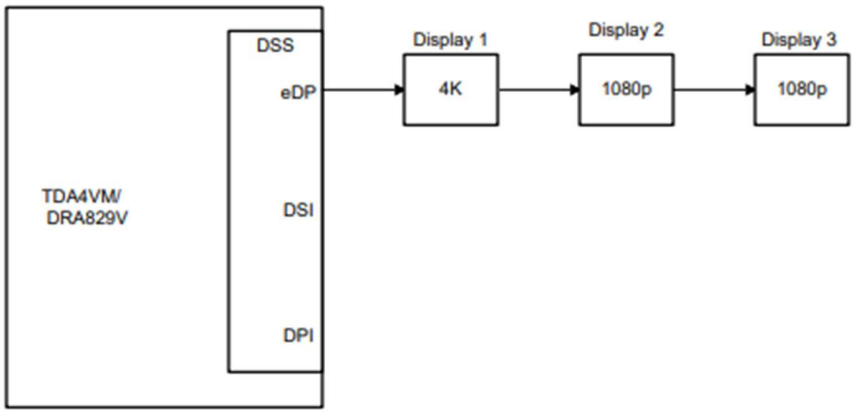


Figure 24 Connecting multiple display via daisy chain

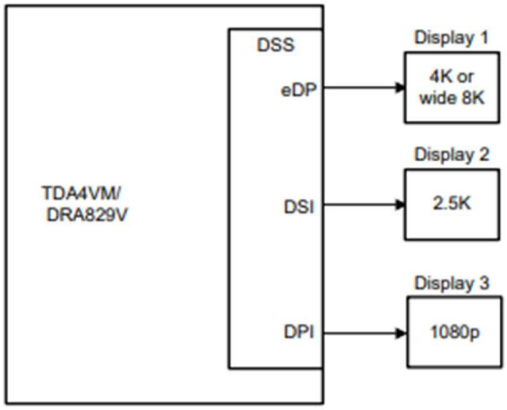


Figure 23 Multiple displays connected in parallel

Camera Subsystem

Similar to the DSS, a camera subsystem can be used to handle the preprocessing and handling of the surround camera system. Taking inspiration from TI's range of products, the TDA4VM/DRA829 camera capture subsystem includes 2x MIPI CSI-2 interface and video processing front end (VPFE) as shown in Figure

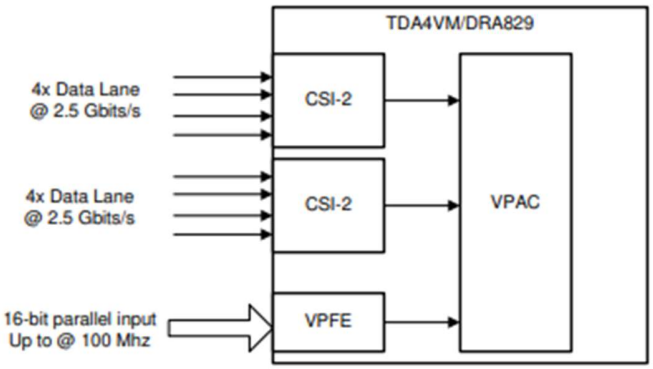


Figure 25 Camera Subsystem architecture

The Vision Pre-processing Accelerator (VPAC) subsystem comprises a set of fundamental vision functions that handle pixel data processing tasks. These tasks include color processing and enhancement, noise filtering, wide dynamic range (WDR) processing, lens distortion correction, pixel remapping for de-warping, on-the-fly scale generation, and on-the-fly pyramid generation. The VPAC efficiently manages these routine tasks, freeing up the main SoC processors (such as ARM and DSP) for the execution of more sophisticated high-level algorithms. Operating in time-multiplexing mode, the VPAC is designed to support multiple cameras. It serves as the front end of the vision processing pipeline, allowing subsequent processing by other vision accelerators or processor cores within the SoC.

Introducing Edge Computing

EC is the paradigm which aims to perform computation tasks at the edge of the network on downstream data on behalf of cloud services and upstream data on behalf of IoT services. The edge architecture is composed of three main layers which are IoT devices, edge, and cloud. AR applications run on Edge devices can respond faster to user input, making the AR experience more natural and immersive. In addition to reducing costs and improving the user experience, Edge computing also enhances the security of AR applications.

AR applications run on Edge devices can respond faster to user input, making the AR experience more natural and immersive. In addition to reducing costs and improving the user experience, Edge computing also enhances the security of AR applications.

Edge based architectures are implemented for mobile devices already. Autonomous vehicles can benefit a lot by utilizing edge based computation for augmented reality computations. A very apt example is EdgeSLAM where the SLAM algorithm is split between the mobile device and the edge device. The architecture is shown in figure as reference.

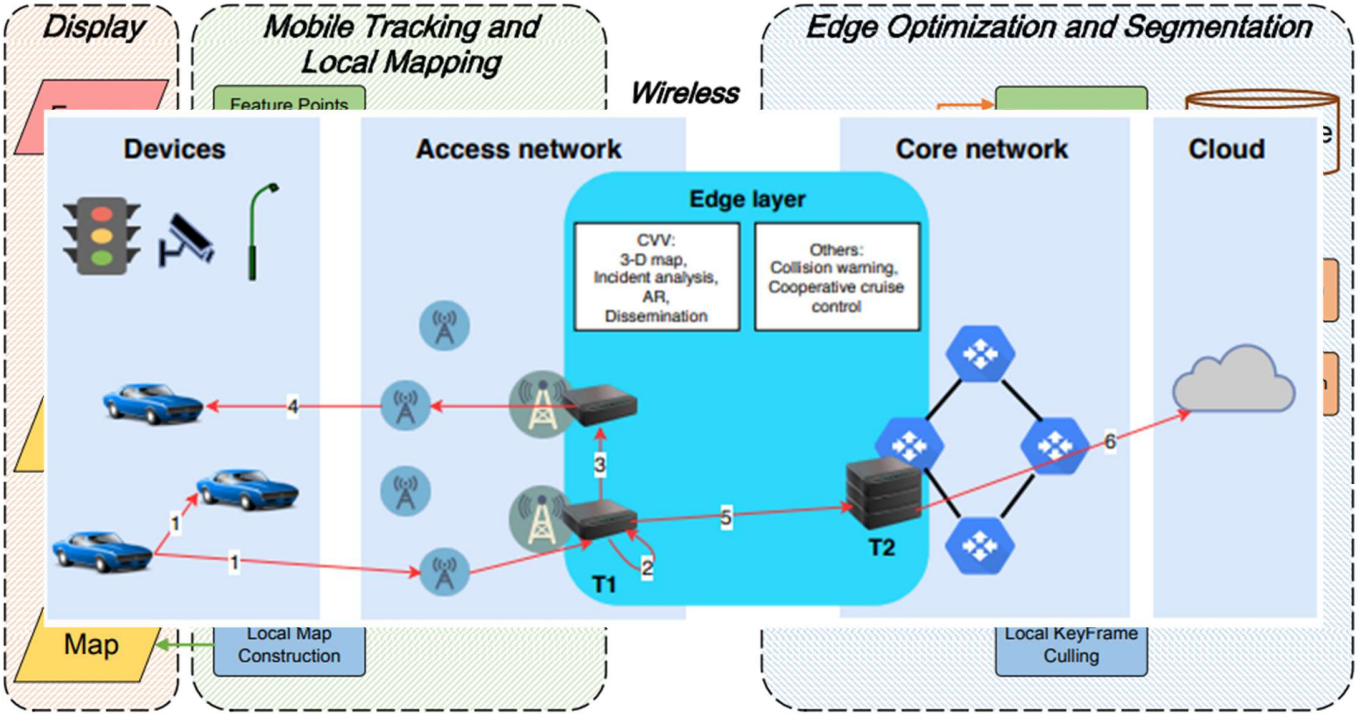


Figure 26 Architecture of EdgeSLAM

By keeping the map creation process on the edge server and limiting the mobile device to tracking, the computation requirements of the mobile device is significantly reduced. This shows a significant improvement in the performance of the SLAM algorithm, with lesser power utilization on the mobile device. A similar analogy can be made to its application in our vehicle environment. Local maps of the environment can be kept in edge servers spread across a region as shown in the figure below and discussed in paper (ARVE: Augmented Reality Applications in Vehicle to Edge, n.d.).

- The placement of the edge layer in the network diagram shows a reduction in communication requirements between the vehicle and cloud.
- This would mean a lower communication technology can be used. That is, 5G technology is not required for all communication. Bringing the power and technology needs down.
- Data is delocalised. All maps and user information is not in a singular cloud making it susceptible to attack.

Why not use Cloud?

As of November 15, 2023, Cruise is pausing all public road operations for its autonomous vehicles (AVs). This includes both supervised and manual operations. Cruise is taking this step to rebuild public trust while undergoing a full safety review. This comes as no surprise to active followers because reported as recently as August 14th of this year (GM's Cruise Falls Down During Concert Cell Overload, n.d.) where as many as 10 Cruise vehicles stopped dead and blocked traffic in the busy North Beach area. Cruise's only response was to state that "A large festival posed wireless bandwidth constraints causing delayed connectivity to our vehicles. We are actively investigating and working on solutions to prevent this from happening again. We apologize to those who were impacted."
This is not an isolated incident as shown by redditor (Bunch of Cruise cars stuck on Gough by Robin, n.d.):

Figure 27 Cruise Autonomous vehicles facing issues. Reported by bystnders on reddit.

Operational stoppages are not isolated incidents and hence companies such as Waymo tend to favor keeping all driving computation inside the car (Waymo keeps 5G way in the background, n.d.). As a result, they tend to have much more expensive cars which do not require advanced communication technology. It is a matter of safety and operation.

These issues can be addressed by utilizing specialized edge servers maintained by the company.

- These servers can be spread across regions and handle location specific data such as maps and user information.

- It can hold AR specific data such as point maps, location specific maps such as shop locations, landmark information etc. For example, as shown in the figure. Louis Vuitton ran an AR based campaign at famous landmarks.

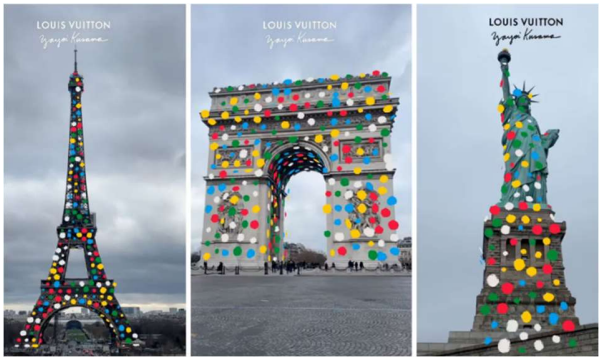


Figure 29 Louis Vuitton Ad campaign utilizing AR

Figure 28 Utilizing AR as informative sources for travel

By offloading only auxiliary functions onto edge servers we can maintain the safety of the user, improve data security and offer advanced features in the car.

VPI

Having provided the Architecture for Different aspects of our Processing, how are we going to integrate these? The answer to this question is the **Vision Processing Interface**

In the hardware theater, VPI operates as a versatile integrator, deploying chiplet clusters strategically. PVA is a power-efficient chiplet, excels in image processing, while VIC and OFA add specialized functionalities. The figure below demonstrates the task distribution of Different Components of our computing system in Stereo Image Processing.

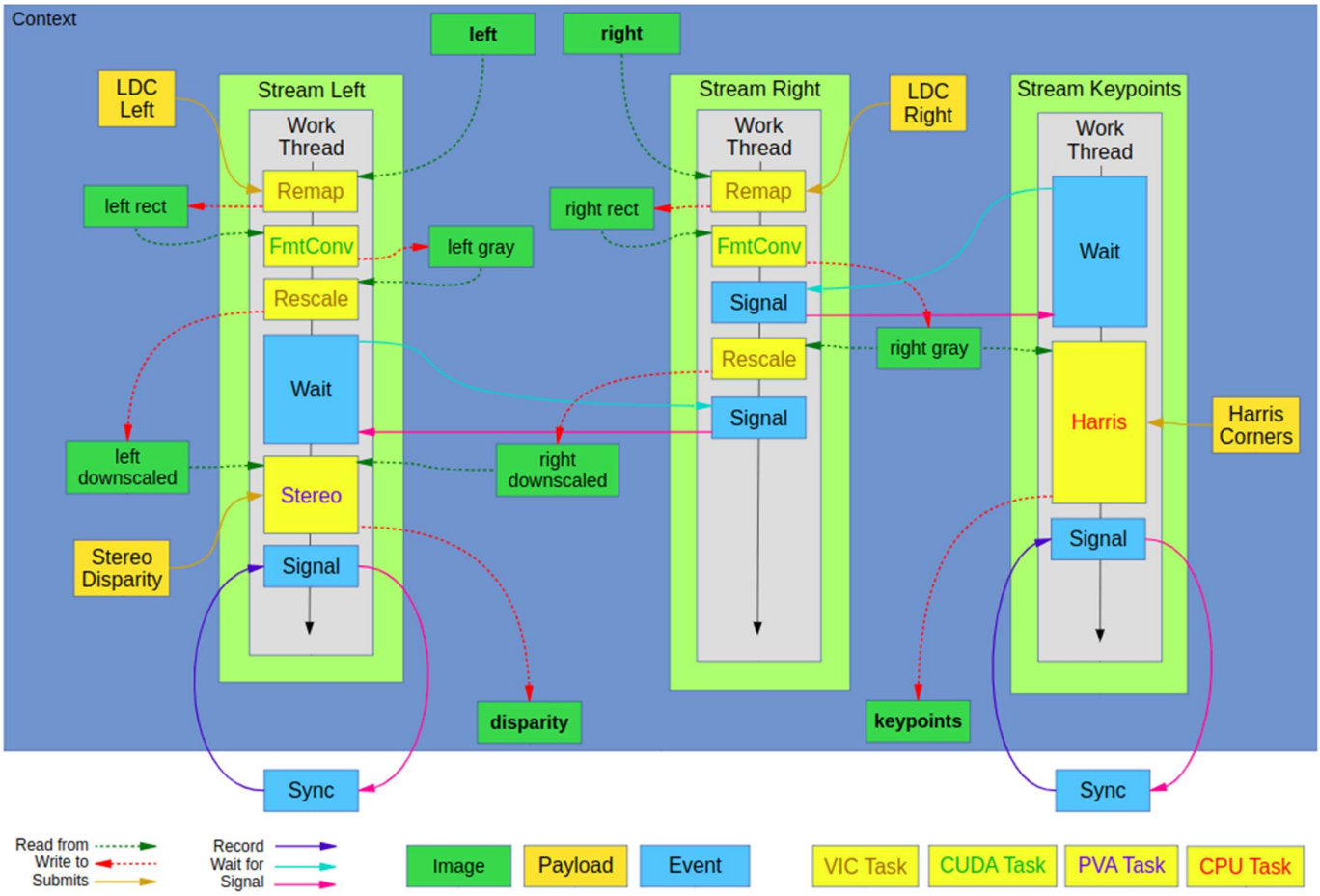


Figure 30 Task distribution of different components of our computing system in Stereo Image Processing.

VIC:	Does stereo pair rectification and downscaling.
CUDA:	Does image format conversion.
PVA:	Does stereo disparity calculation.
CPU:	Handles some preprocessing and extraction of Harris corners.

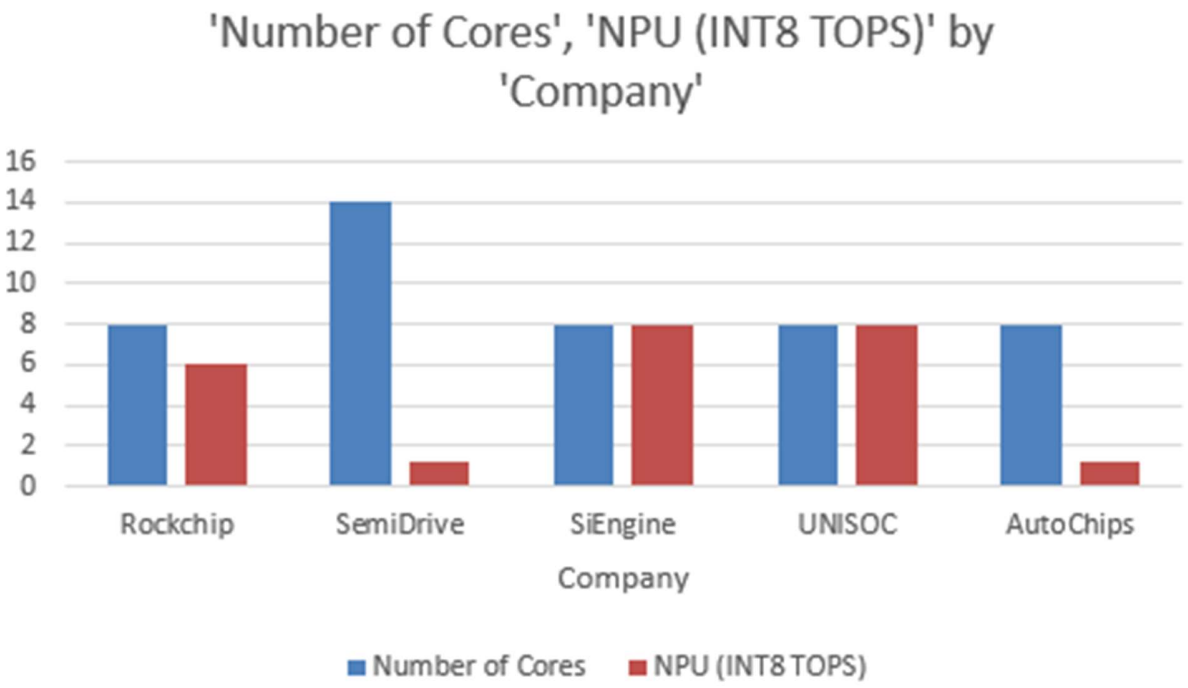
Figure 31 Task Distribution of Various Components of our computing system in SIP

AN EYE ON CHINESE COMPETITION AND SOURCING OF PARTS

China at once was the largest market for JLR, the Chinese companies also threaten to undercut many OEM Manufacturers. Hence, a close look at Chinese solutions and their analysis is critical for JLR to compete with them in both China and other Competing Markets.

Company	Rockchip	SemiDrive	SiEngine	UNISOC	AutoChips
Model	RK3588M	X9U	Dragon Eagle One	A7870	AC8025H
Process	5nm	16nm	7nm	6nm	-
Number of Cores	8	14	8	8	8
CPU-Architecture	4xA76@2.2Gh4xA55@1.7Ghz	14xA55	4x A76 @ 2.4GHz;4x ASS	1x A76 @ 2.7GHz; 3x A A76 @ 2.3GHz;4x A55@ 2.0GHz	6*A55+2*A76+2*R5F
CPU Main Frequency	2.2Ghz	2.0Ghz	2.4Ghz	2.7Ghz	-
CPU Compute DMIPS)	100K	100K	90k	93K	60K
GPU ARCHITECTURE	Mali G610 MC4	-	-		
GPU Compute GFLOPS)	512	300	900		120
NPU (INT8 TOPS)	6	1.2	8	8	1.2
Automotive Grade	AEC-Q100	AEC-Q100 ASIL-B	AEC-Q100	AEC-Q100	AEC-Q100
Planned SOP	2023	2021	2023	2022	2023

Figure 32 Study of Chinese Market



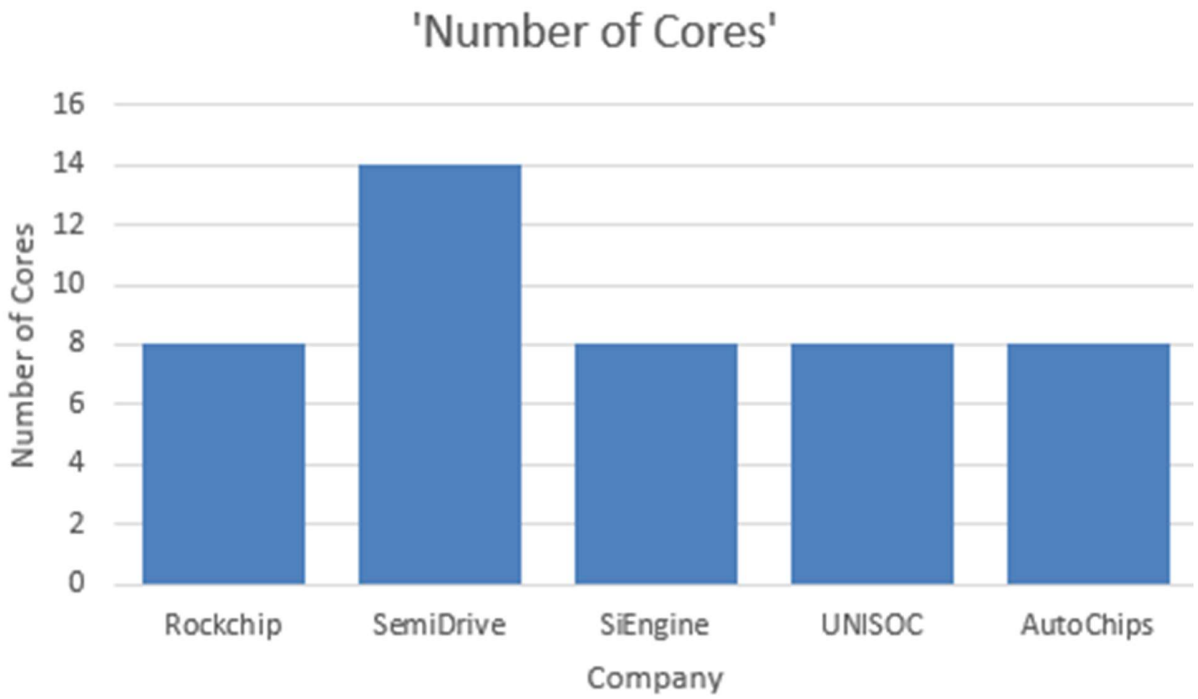


Figure 33 Comaprison among various companies by number of cores

COMPARING CHINESE COMPANIES AGAINST OUR PROSPECTIVE SUPPLIERS AND HEAVYWEIGHTS

Feature	NVIDIA Orin	Snapdragon Drive	Rockchip RK3588M	UniSoC A7870
GPU	Ampere architecture, 2048 CUDA cores	Adreno™ 320 GPU	Mali-G610 MC4	PowerVR G6610
CPU	12-core Arm Cortex-A78AE v8.2 64-bit CPU	1.5GHz Quad-Core Krait™ CPU	8x A76 @ 2.2GHz + 4x A55 @ 1.7GHz	6x A76 @ 2.7GHz + 2x A75 @ 2.2GHz
DL Accelerator	2x NVDLA v2.0	Hexagon™ QDSP6 DSP	N/A	NPU (INT8 TOPS) 6
Vision Accelerator	PVA v2.0	N/A	N/A	N/A
Memory	32GB 256-bit LPDDR5, 204.8 GB/s	4GB (3GB accessible)/2GB DDR3L-1066	8GB LPDDR4x	8GB LPDDR4x
Storage	64GB eMMC 5.1	64GB eMMC	128GB eMMC 5.1	128GB eMMC 5.1
Camera Connector	16 lane MIPI CSI-2 connector	FPD-LINK II Input	2x MIPI CSI-2 2.0	4x MIPI CSI-2 3.0
Display	DisplayPort 1.4a (+MST)	LVDS to FPDLINK III	4K60Hz HDMI 2.0	4K60Hz HDMI 2.0
MicroSD Slot	UHS-1 cards up to SDR104 mode	One SD card slot	One SD card slot	One SD card slot
Chinese Manufacturer	No	Yes	Yes	Yes
Overall Characterization	High-performance and power-efficient for demanding applications.	Budget-friendly option for basic functionality.	Mid-range option offering a good balance of performance and cost.	Another mid-range option with decent performance and features.

Table: Comparison of possible suppliers to Chinese Comapnies

Qualcomm was the largest installed SoC provider in Chinese cars of 2022, 43.6% to be precise.

Here is a brief analysis of Hardware suppliers for our competitors

Hardware Supplier	Automobile Companies Using Their Hardware
NXP	Tesla, BMW, Mercedes-Benz, Volkswagen, Audi, GM, Ford, Volvo, Toyota, Hyundai, Tata, Great Wall Motor, GAC, Changan, SAIC, Geely, BAIC, FAW Hongqi, Chery, Dongfeng Voyah
Texas Instruments (TI)	Tesla, BMW, Mercedes-Benz, Volkswagen, Audi, GM, Ford, Volvo, Toyota, Hyundai, Nissan Renault
Renesas	Tesla, BMW, Mercedes-Benz, Volkswagen, Audi, GM, Ford, Volvo, Toyota, Hyundai, Honda, Jaguar Land Rover
Qualcomm	Tesla, BMW, Volkswagen, Audi, GM, Ford, Volvo, Toyota, Hyundai, Honda, Nissan Renault
Intel	Tesla, BMW, Mercedes-Benz, Volkswagen, Audi, GM, Ford, Volvo, Toyota, Hyundai, Honda, Nissan Renault, Jaguar Land Rover
Samsung	Tesla, BMW, Mercedes-Benz, Volkswagen, Audi, GM, Ford, Volvo
NVIDIA	Tesla, BMW, Mercedes-Benz, Volkswagen, Audi, GM, Ford, Volvo, Toyota
Telechips	Hyundai, Kia
AMD	Tesla, BMW, Mercedes-Benz, Volkswagen, Audi, GM, Ford, Volvo
Rockchip	Great Wall Motor, BYD, Geely, Changan, SAIC, FAW Hongqi, Chery, Dongfeng Voyah, Li Auto, NIO, Xpeng, WM Motor, Hozon, Human Horizons, Leap motor
SemiDrive	Great Wall Motor, SAIC, Li Auto, NIO, WM Motor, Hozon, Human Horizons
MediaTek	Great Wall Motor, BYD, Geely, Changan, SAIC, BAIC, Chery, Dongfeng Voyah
AutoChips	NIO
SiEngine Technology	NIO
Huawei Hisilicon	Huawei
UNISOC	GAC, Changan, SAIC, Geely, BAIC, FAW Hongqi, Chery, Dongfeng Voyah, Li Auto, NIO, Xpeng, WM Motor, Hozon, Human Horizons, Leapmotor
AllwinnerTechnology	BYD, Geely, Changan, SAIC, BAIC, Chery, Dongfeng Voyah

Table: Supply for competitors across the globe

Here is a table summarizing the various prospective supplier for Jaguar Land Rover and their current progress/feasibility with the individual parts concerned

Feature	Nvidia	AMD	Qualcomm	MediaTek	Renesas
CPU	Off-the-shelf or custom ARM	Ryzen Automotive SoCs	Snapdragon Automotive SoCs	Custom automotive SoCs	Automotive SoCs
AI Accelerator	Drive AGX platform with GPUs and DLAs	RDNA 2 architecture GPUs	Hexagon DSP and custom DLAs	APU (AI Processing Unit)	Dedicated AI accelerators
Programmable Vision Accelerator	Have developed V2 of its programmable vision architecture	N/A	Adreno Visual Processing Unit (VPU)	Custom vision processing units	Vision processing units
VIC 2D Engine	VIC 2D Engine	N/A	Integrated video codec engine	Integrated video processing engine	Integrated video processing engine
GPU	Drive AGX platform with GPUs	RDNA 2 architecture GPUs	Adreno GPUs	PowerVR GPUs	Various GPU options
GPCs, TPCs, SMs	Full utilization	Partial utilization depending on SoC	Partial utilization	Partial utilization	Partial utilization
CUDA Cores	Yes	Yes, limited compared to Nvidia	Limited support	Limited support	Limited support
Tensor Cores	Yes	No	Custom for specific applications	Custom for specific applications	Custom for specific applications
Memory	LPDDR4x and HBM2	LPDDR4x and LPDDR5	LPDDR4x and LPDDR5	LPDDR4x and LPDDR5	LPDDR4x and LPDDR5
Video Codecs	NVDEC and NVJPEG	Video processing capabilities integrated into Ryzen SoCs	Custom video codecs for specific applications	Custom video codecs for specific applications	Custom video codecs for specific applications
Partnerships	Strong partnerships with automotive manufacturers	Collaborations with automotive manufacturers	Strong partnerships with automotive manufacturers	Collaborate with automotive manufacturers	Collaborate with automotive manufacturers
Sources	https://www.nvidia.com/en-us/self-driving-cars/, https://developer.nvidia.com/cuda-gpus, https://www.nvidia.com/en-us/data-center/tensor-cores/	https://www.amd.com/en/products/embedded-ryzen-series, https://finance.yahoo.com/quote/AMD/	https://www.qualcomm.com/products/automotive/qualcomm-automotive-solutions-ecosystem-program, https://developer.qualcomm.com/software/adreno-gpu-sdk, https://en.wikipedia.org/wiki/Qualcomm_Hexagon	https://www.mediatek.com/products/automotive, https://www.mediatek.com/technology/artificial-intelligence	https://www.renesas.com/us/en/products/automotive-products, https://www.renesas.com/us/en/products/automotive-products/automotive-system-chips-socs

Table: Part Offerings in Market

Apart from these parts there are cutting Edge Analog Processors and Neural Processors that we

	TSMC	Intel	TSMC	CEA-Leti	Intel	Intel
Product Name	CoWoS	EMIB	InFO	INTACT	Foveros	Co-EMIB
Integrated type	2.5D	2.5D	3D	3D	3D	3D
Interposer type	Passive	Passive	-	Active	Active	Active
Interconnect pitch (µm)	40	55	-	20	36	36
PIM Application	NVIDIA GP100	Agilex FPGA	Apple A10 processor	-	Lakefield processor	Ponte Vecchio GPU
Bandwidth	717 GB/s	896 GB/s	-	527 GB/s	-	2 Tb/s
Power	235 W	-	-	30 W	7 W	600 W
Frequency (GHz)	1.4	1.5	-	1.15	1	1.37
Latency	-	~60 ps	-	0.6 ns/mm	-	-
Yield	High	High	High	High	High	High
Reusability	High	High	High	High	High	High
Application	HPC Edge Computing	Data Center Networking Edge Computing	Mobile IoT	HPC AI Edge Computing	Mobile PC	Date Center Machine Learning HPC

Table : Popular architectures and their Implementations

AUTOMOTIVE CYBER SECURITY

List of Key Players in Automotive Cyber Security Market

- Intel Corporation
- Advanced Micro Devices
- NXP Semiconductors
- Cisco Systems
- Harman International Industry
- Escrypt Embedded System
- Secunet AG
- Argus Cyber Security
- NNG Software Developing and Commercial LLC.

Major Standards Defining Automotive Grade

Fig:

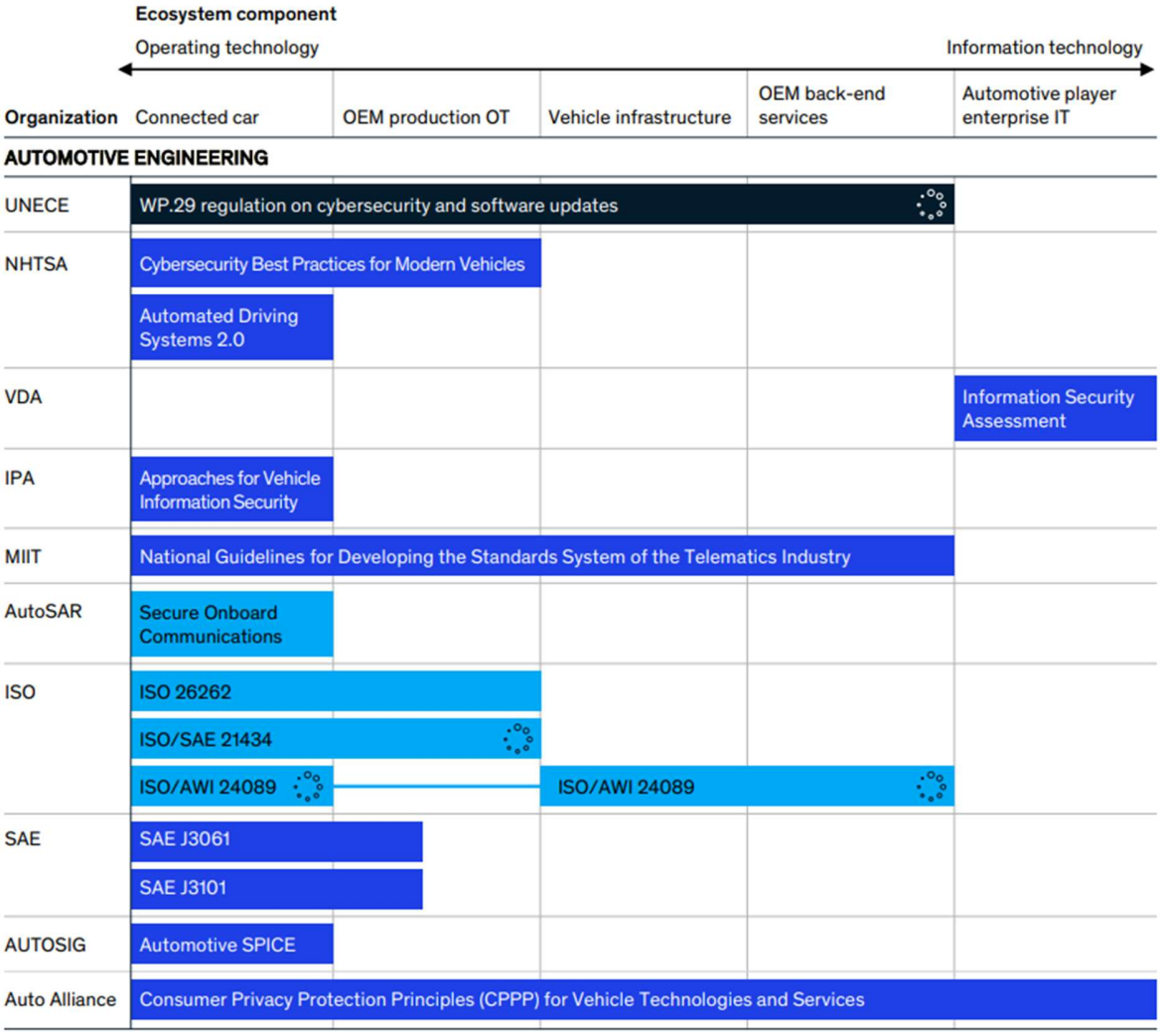


Figure 34 Different Standards for Cyber Security

- ISO-26262: It puts up functional safety requirements for automotive and is internationally recognized.
- ISO-21434: Framework that standardizes Automotive Cyber Security requirements
- EVITA: It is a project for cyber security in automotive, i.e. to to verify, design and prototype architecture and on-board networks in automotive to protect security-specific components against tampering, and protection of Vehicle-to-X communication.
- AUTOSAR: AUTomotive Open System ARchitecture is a partnership of stakeholders of automobile industry aiming to create standardization in software architecture for ECUs.

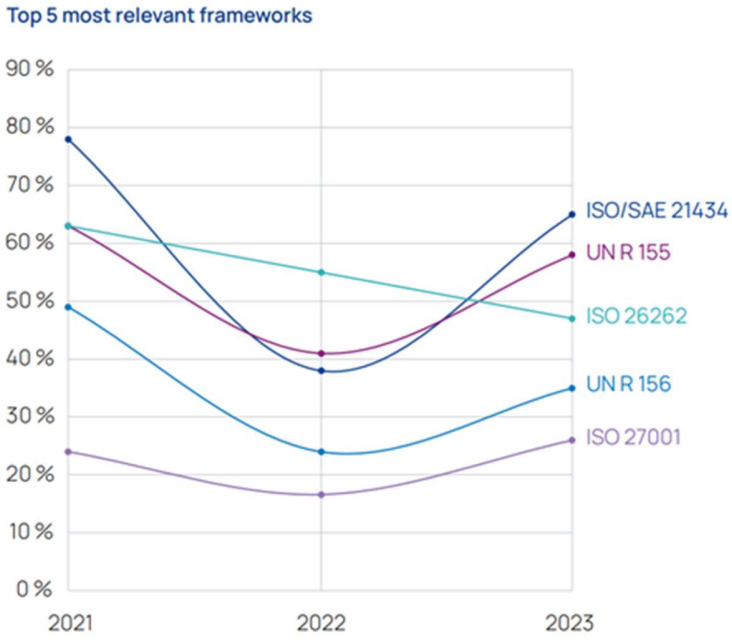


Figure 35 : Implementation of adequate security by mature automotives (survey data)

The above is the graph from a survey report of ETAS depicting the utilization of various standards by major automotives. This indicates a growth of standards utilization.

The goal of International Standard, ISO/SAE 21434 is:

(a) To put up requirements of cyber security and risk management
(b) To put up a framework to manage these requirements without needing to indicate technologies, rather give a reference, which is useful for legal aspects too.

Here are some statistical representations of cyber attacks in Automotive.

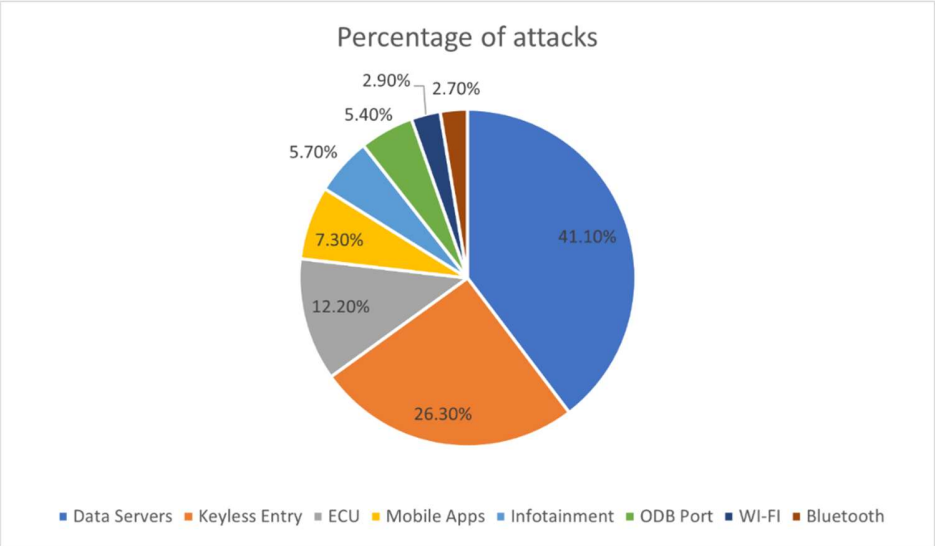


Figure 36 Percentage of various attacks in Automotive

ECU is one of the most in-vehicle vulnerable sites because it operates with CAN buses that do not go encrypted. Insecure CAN connectivity poses a threat to secure communication inside automotive, as the CAN protocol does not go through cryptographic data exchange. The system, therefore, needs an intrusion detection system to monitor adversaries.

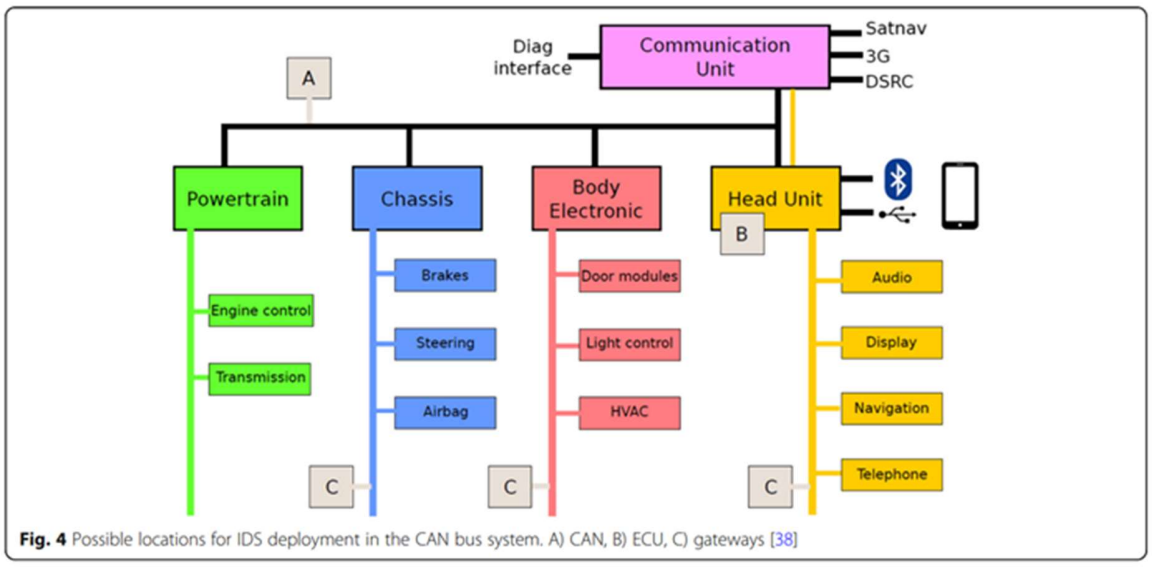


Figure 37 Possible points for intrusion detection system Deployment

The above picture shows possible entry points or loop holes for a successful attack. It thus needs an intrusion detection system in place for the safe operation of the concerned subject.
The task of an Intrusion detection system is to identify attacks on vehicles and report it to the Vehicle Security Operation Centre.
For the solution to all the communication needs at broad level(not intra-chiplet), we could suggest **ETAS Escrypt Cycur Family**.

ETAS Escrypt CycurX

ETAS Escrypt Cycur family offers automotive security at all levels. CycurHSM offers secure encapsulation of all the security mechanism in a determined software component which is kept seperated from the OS and application software

Features and Overview-

(a) Safety Qualification (ISO 26262, ASIL-D)
(b) It has a modular structure
(c) Full compatability with AUTOSAR, SHE, SHE+
(d) It has full support of HSM technology (Infeneon, NXP, STmicroelectronics, Renesas)

It offers security in:

(a) Secure Booting of the ECU
(b) Secure in-vehicle communication
(c) ECU Component Protection
(d) Secure debug, storage and log

Particularly for the vast network of communication through CAN, the Escrypt CycurECU has the following:

- It monitors forwarded CAN traffic & detects potential attacks or adversaries.
- Report is attack either locally or to Vehicle SOC
- Detection of ECU based on heuristic and signature.
- It is totally consistent with the generic IDS manager concept.
- It works as a “smart sensor” to figure out intrusions on CAN.

Example Features in Detection:

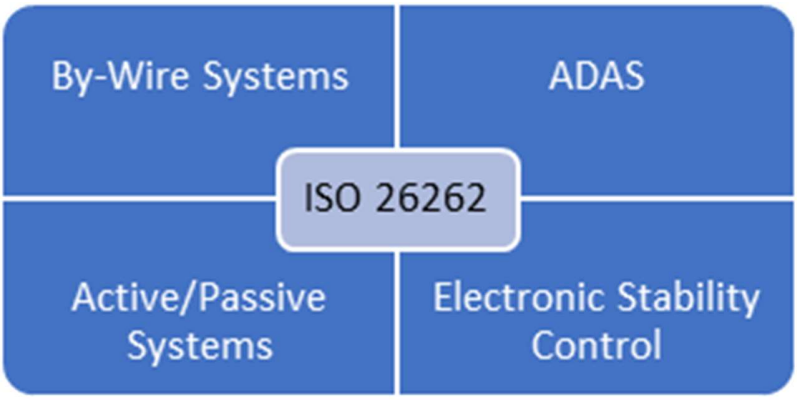
- Enables observing frequency of message to detect fault injection.
- Allows comparison of all messages on the buses and whitelist to detect false messages.
- Detection of malicious diagnostic requests while driving, e.g. detect any kind of attempt to shut down an ECU.

Safety Architecture Verification

Since semiconductor chips are open for random failures, there have been developed Design for Test (DFT), techniques , for example, scan of full design, and then generation of automatic test pattern to stimulate a fault. This fault can then be propagated to an observable output pin. For automotive verification of safety, the approach is as:

- Injection of a random fault
- Propagation of that fault
- Checking of the safety mechanisms catching the fault
- Classification the fault
- Safety metrics generation

ISO 26262 Covers the following.



The automotive safety integrity levels here are based on 3 variables.

How does this affect our chiplet design and intelligent hardware?
Any tool hardware of software must qualify under the ISO 26262 standards.
Each tool is analyzed to supply a software confidence level (TCL). The TCL and ASIL decide the level of reliability or confidence we need in a software tool.

Figure 38 Increasing order of tool confidence levels

Equations for each:

$$\textit{Single Point Fault Metric (SPFM)} = 1 - \frac{\Sigma(\lambda SPF + \lambda RF)}{\Sigma\lambda} = \frac{\Sigma(\lambda MPF + \lambda S)}{\Sigma\lambda}$$

$$\textit{Latent Fault Metric (LFM)} = 1 - \frac{\Sigma(\lambda MPF,L)}{\Sigma(\lambda - \lambda RF - \lambda SPF)} = \frac{\Sigma(\lambda MPF,D,P + \lambda S)}{\Sigma(\lambda - \lambda RF - \lambda SPF)}$$

$$\textit{Probabalistic Metric for Hardware Failure (PMHF)} \sim \lambda SPF + \lambda RF + (\lambda MPF,D * \lambda MPF,L * T_{\text{Lifetime}})$$

Figure 39 Safety metrics calculations

The five FMEDA safety metrics are:

1. FIT- Failure In Time
2. SPFM- Single Point Fault Metric
3. LFM- Latent Fault Metric
4. PMHF- Probabilistic Metric for Hardware Failure
5. DC- Diagnostic Coverage

The chiplet being manufactured must have to pass the minimum requirement put in the standards.

Fault Mode	Fault Type	Observed time	Detected Time
Soc_top.u_proc.u_alu.ma[0]	SA1	10200	
Soc_top.u_proc.u_alu.ma[1]	SA0	14700	
Soc_top.u_proc.u_alu.ma[2]	SA1	10200	11400
Soc_top.u_proc.u_dma.n13	SA1	10200	128400
Soc_top.u_proc.u_dma.1429	SA1	18900	20400
Soc_top.u_ctrlmon.u_jtag.cmd[0]	SA0	11700	16200
Soc_top.u_hw_accl.u_decomp.ctrl[0].ma[0]	SA0	14700	
Soc_top.u_hw_accl.u_decomp.ctrl[0].ma[1]	SA1	10500	128400
Soc_top.u_hw_accl.u_decomp.ctrl[0].ma[2]	SA1	10500	128400
Soc_top.u_ctrlmon.u_jtag.cmd[0]	SA1	21900	
Soc_top.u_periph.u_spl_top.n1431	SA1	10200	243600
Soc_top.u_periph.u_spl_top.n1431	SA1	10200	243600

Part/Sub-Part	SM	FIT	SPFM	LFM	PMHF
uP	CPUHang	1.24	98%	99%	0.23
DMA	Out of order xfer	0.45	90%	85%	0.08
SRAM	Computer memory	0.94	99%	99%	0.13
PCIe	Computed data	0.65	92%	90%	0.009

**Blocks in red are detected faults*

Figure 40 Typical Fault Classification/FMED Analysis from Safety Mechanism

Safety Island

Functional safety requirements that are devised by ISO 26262 is to negate the failure of any safety-critical system

Safety island is to manage and control the safety content inside an SoC by:

- Signaling the failures
- Enabling their recovery
- Adaptation with the needs of future

These are some added functional safety block(in gray colour) in a typical automotive SOC.

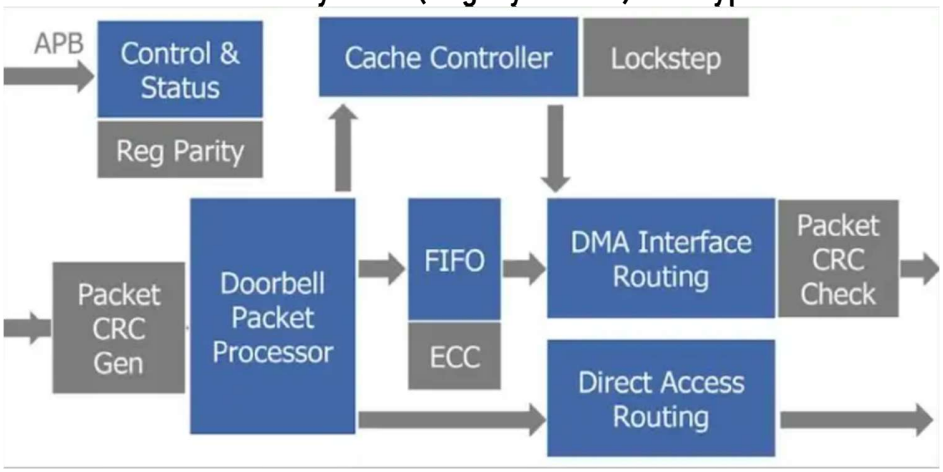


Figure 41 Block for functional safety (Gray Colored)

Chip-Level Security

In a chiplet-based system, there is a possibility of untrusted chiplet in the supply chain. This can lead to:

1. Risk for IP Piracy/Modification/Unauthorized Access
2. Untrusted chiplets can perform various Cyber-attacks, such as:
    - Side-Channel Attacks: Untrusted chiplets could potentially leak sensitive information through side-channel attacks, compromising the confidentiality of data processed by the system.
    - DOS Attack: Transmitting huge number of redundant packets across chiplets can stop the service of other chiplets.

All chiplets therefore must be sourced from authenticate and trusted suppliers

For secure communication between chiplets, either each chiplets can have their own security algorithm or a centralized chiplet can take charge of the overall security.

We suggest for a Central Security for the following reasons:

1. Many redundant security features is minimized causing better resource utilization.
2. The communication system inside the system is hard fixed, that suits good for a central security policy.
3. A decentralized security system needs high customization and extensive design research, and the current market is more consistent with central security solution.

A centralized security system in our regards refers to a hardware security Module.
In the case of an SoC, it is impossible to monitor internal signals on a 5nm chip; monitoring on the interposer in a multi-chiplet design is much more feasible. Therefore, the communication between the chiplets of anything security-related needs to be encrypted.

'When you break it down into chiplets, the SiP is only as good as the least secure chiplet.'

-Scott Best (Rambus Security)

For secure inter chip communication, the Hardware Security Module performs a sequence of operation. Advanced Encryption Standard (AES) is employed for symmetric-key encryption, which ensures the confidentiality of sensitive data within the memory. Hash-based Message Authentication Code (HMAC) verifies message integrity and authenticity, safeguarding against tampering or unauthorized modifications. Rivest–Shamir–Adleman (RSA) algorithm is used for public-key cryptography, supporting secure communication, key exchange, and digital signatures. Elliptic Curve Cryptography (ECC) is leveraged for efficient public-key operations with shorter key lengths, particularly beneficial in resource-constrained automotive environments. Additionally, a True Random Number Generator (TRNG) is essential for providing a reliable source of unpredictability, contributing to the generation of secure cryptographic keys. The key is stored inside a memory unit inside the HSM.

Apart from this, security treatment also comes from a hypervisor. A Hypervisor is a Virtual Machine Monitor (VMM), which is performs virtualization. It allows multiple-operating systems running on at the same time on a single physical machine, and therefore provides isolation between virtual machines (VMs). Hypervisors create isolated environments for each virtual machine. This isolation helps prevent security breaches in one VM from affecting others. Even if a security vulnerability is exploited in one VM, the others remain unaffected. Another functional safety unit in the system is Lockstep. It is a functional safety requirement where multiple processors runs in parallel to ensure the performance of individual processors, and look for single point fault.

Hardware Security Module

An HSM(Hardware Security Module) is a hardware specifically designed fpr providing secure key management and tamper-resistant cryptographic operations. HSMs play a crucial role in securing sensitive data and transactions, offering a level of protection that is challenging to achieve with software-only solutions. It takes charge of all the cryptographic operations. It contains the Physical Unclonable Function and is designed to be tamper resistant.

Concerns with Hardware Security Module

- APIs are not well standardized, so interfacing between different systems is more complex.
- Low flexibility of the functionality provided.

The Hardware Security Module along with the cryptographic operations, Memory, and Secure Boot forms a Trusted Platform Module, that acts as a Root of Trust for the entire system.

For the closest solution that covers all the security need, we could suggest two IPs:

Rambus RT-645

This is an automotive defined co-processor IP and a Root-of-Trust designed to safeguard vehicle from adversaries in V2X communication, ADAS, ECU platform management, infotainment and other critical systems. It is an EVITA (E-Safety Vehicle Intrusion Protected Applications), ISO-26262 and ASIL-D standardized complete Hardware Security Module. This IP supports all major processors i.e. ARM, , x86, RSIC-V, etc.

This HSM gives broad range of security features for automotive chips that includes secure boot, secure debug, secure firmware updates, attestation, device personalization, authentication, key and data provisioning secure feature and configuration management. It has support for cryptographic acceleration, and secure key and data storage.

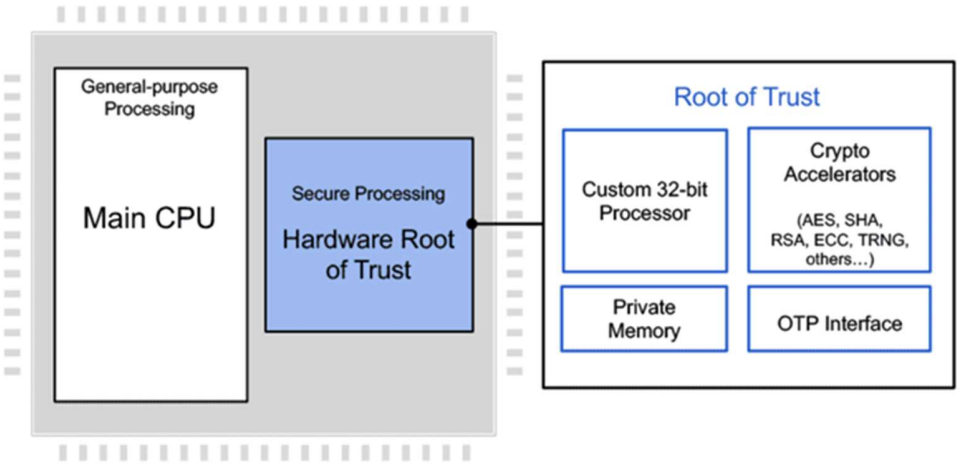


Figure 42 RT-645 HS- RoT Functions

It also comes with an SOA container development kit that allows for custom SOA containerization for specific use cases.

Security Features

1. It has Root of Trust cores that employs a custom secure RISC-V processor
    - RT645- It is ASIL-D Certified
2. Its software and hardware safety mechanisms offers more the standard ASIL-B/ASIL-D safety requirements
3. It has anti-tamper feature and Security in in-core processing
4. Protection for all components in the core by its Multi-layered security

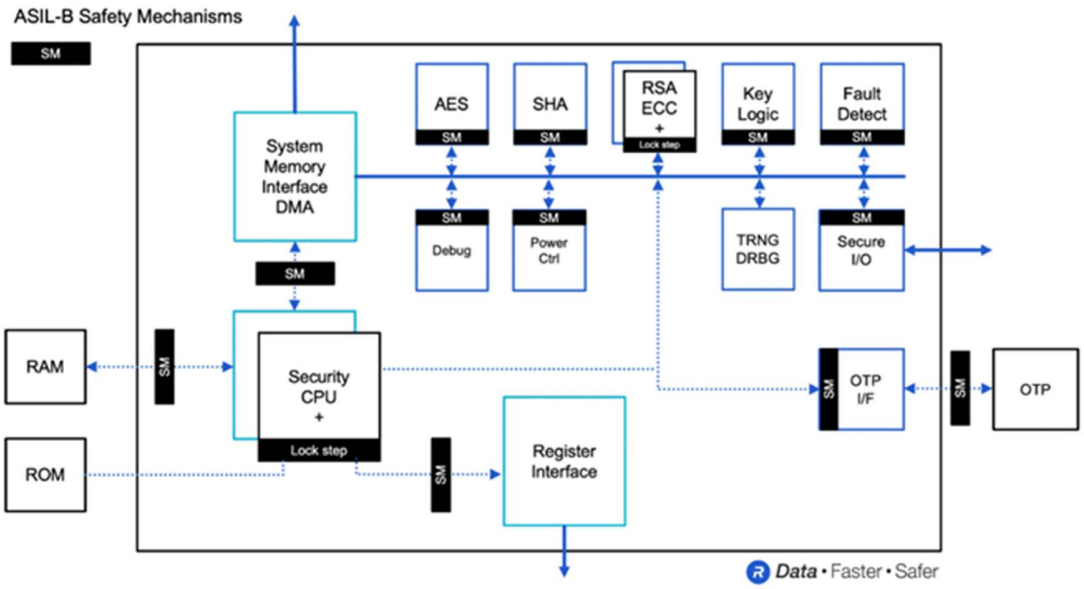


Figure 43 Ransom RT-640 ASIL-B Safety Mechanism Block Diagram

The above is a typical ASIL safety Mechanism Block Diagram that communicates with encrypted data to all other logic units.

Flexibility

- Third-party applications can operate securely within a designated secure boundary, each with individually assigned security permissions.
- Comprehensive development environment enables OEMs and users to effortlessly create secure applications (containers).
- Offers standard application containers for common use cases.
- Enables secure provisioning of keys and firmware during manufacturing or in-field operations.
- Supports the incorporation of multiple roots of trust within a single secure core.

Security Models

- It has hierarchical privilege
- Policy for key management is secure
- Isolation, protection and access-control based on hardware
- Offers secure policy for error management

Cryptographic Accelerators

- It includes RSA, ECC, HMAC, AES, RBG, TRNG (which is configuration-dependent)

Offering form security modules

- It has canary logic that protects against over clocking and glitching.
- Security in key derivation and transport
- Feature-management
- Secure testing and debugging
- Life-cycle management

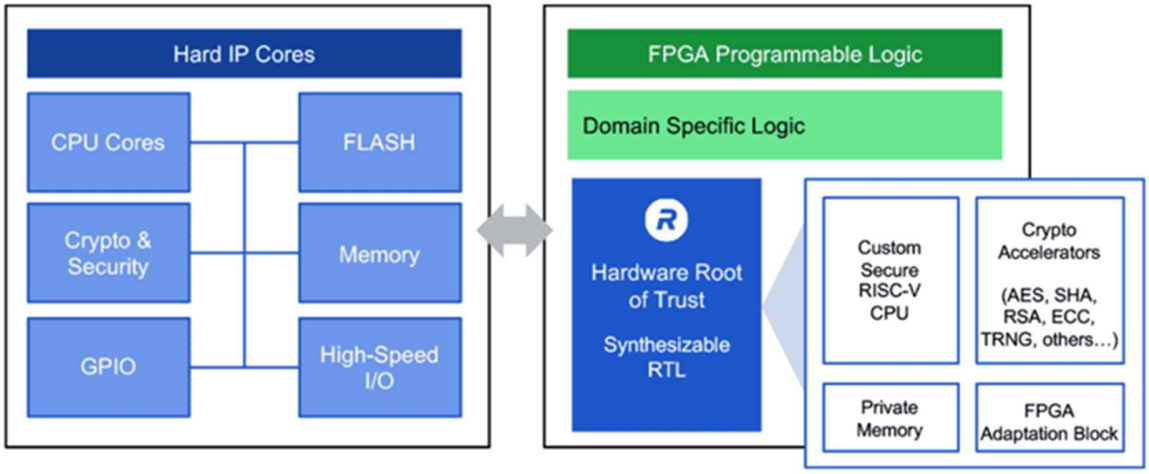


Figure 44 Secure Processing inside FPGA Programmable Logic

Ceva Fortrix SecureD2D IP

This IP offers highly secure die to die communication. This intellectual property (IP) is to communicates with hardware-based crypto using a secure bus fabric and comprises an RISC-V controller. These accelerators execute fast encryption and decryption operations, enabling various cryptographic functions like SHA2, AES, ECDSA, among others. This IP package includes DMA and SRAM controllers, which is an API(a low-level firmware), and a customizable high-level application, offering integration into secure chiplets.

This IP incorporates both follower and leader termination points, that secures chiplet dies sourced from diverse vendors and global supply chains.

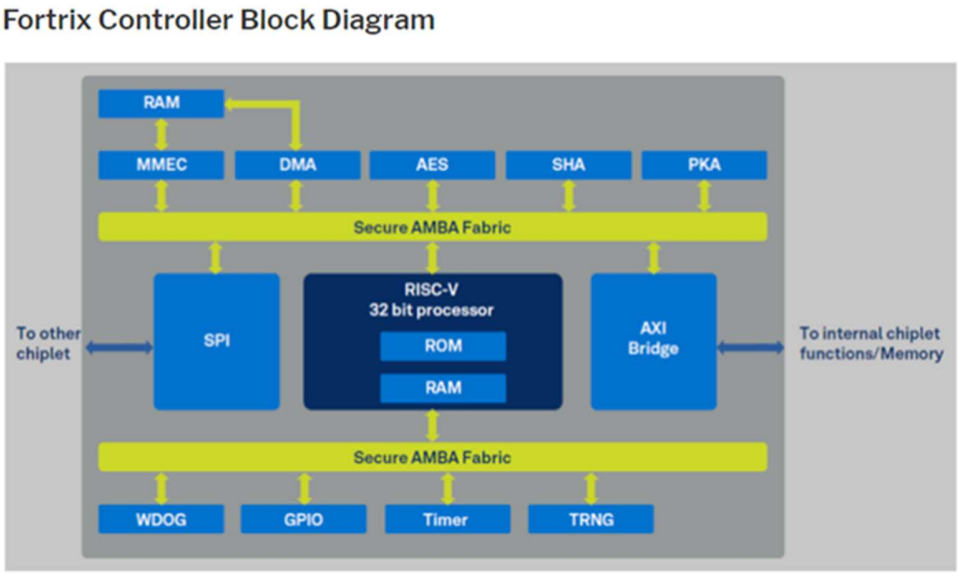


Figure 45 Example- Fortrix Controller Block Diagram

- It protects from many securities threats
- It allows for real time threat protection expansion as threats evolve with time
- Easy implementation into a Heterogeneous system
- Its crypto engines are very efficient for power consumption and compute cycles

Main hardware-blocks from Fortrix are:

- RISC-V processor
- Secure bus fabric
- Control for DMA
- Control for SRAM
- Control for AES encryption/decryption
- Control for SHA256
- Public Key Accelerator (PKA) w/ ECDSA 256/384
- SPI controller/ flash support
- True Random Number Generator (TRNG)

Chiplet Communication and Latency

Developing the interconnect for chiplets requires substantial research and development (R&D) efforts. This includes navigating longer signal routes that may lead to increased impedances, reduced available bandwidth, higher power consumption, and heightened latency. Chiplets employ multiple chips to perform the same functions as a single, monolithic processor. However, this choice extends communication distances between chiplets, increasing latency levels.

Chiplet optimisation

Core level optimization:

Before assembly, chiplets are tested, but inherent parametric variations within each chiplet result in potential differences between cores, especially with higher core counts. AMD has observed up to 200MHz variations in maximum frequency(Fmax) among cores. An algorithm at boot can be used to characterize cores, generating a Fmax-ordered list for the OS. This list guides optimal scheduling, directing high-performance threads to the fastest core. A similar process identifies the best-performing core group for multithreaded workloads. These characterizations occur at every boot, adapting to changes over time, ensuring sustained high performance by selecting the fastest core under current circumstances.

Cache level optimization:

One strategy to reduce latency in chiplet-based processors is adjusting the last-level cache size. This helps offset potential latency differences caused by the physical distance between chiplets and the memory controller, ensuring competitive performance with monolithic designs. Incorporating crosslinks across memory will allow for faster access to cache levels.

Interconnect level optimization:

PHYSICAL (PHY) level

	112G USR/XSR	HBI, BoW, Proprietary	Design Impact
Architecture	Serial	Parallel	Use/Package
Data rate	20 to 112Gbps	1 to 8Gbps	Bandwidth
Pins	10's	1000's	IO Limited
Power	1.5 to 2.0 pJ/bit	0.5 to 1.0 pJ/bit	Link Energy
Latency	High	Low	Performance
BER	FEC	FEC (optional)	Reliability
Package	Substrate	Interposer	Cost/Complexity
Channel Loss	-8 to -12 dB	-4 to -8 dB	Reach
Standardization	IEEE	Mixed	Interoperability

To minimize latency in chip-to-chip communication, it is essential to evaluate the trade-offs between parallel interfaces (such as AIB) and serial interfaces (like SerDes). While parallel interfaces provide low latency, they may require a silicon interposer, increasing overall costs. On the other hand, SerDes offers a more straightforward solution for short-distance communication within the same package. Striking the right balance between these options is crucial for optimizing performance and cost-effectiveness in chip-to-chip communication systems.

Parallel	Serial
Multiple Connections	Single connection
Susceptible to Em interface	Robust EM performance
Consume more power	Saves power
Bigger ICs with complex packages	Fewer pins make IC compact
Practically no latency	Add latency
Challenging skew balancing requirements	Clock can be recovered from the data

Serial Interfaces

A Serializer/Deserializer (SerDes) is a functional component that facilitates the serialization and deserialization of digital data. It is specifically employed in high-speed chip-to-chip communication. SerDes implementations are utilized in contemporary System-on-Chips (SoCs) designed for applications such as high-performance computing (HPC), artificial intelligence (AI), automotive, mobile, and Internet of Things (IoT). These implementations can support diverse data

rates and standards such as PCI Express (PCIe), MIPI, Ethernet, USB, USR, and XSR. A typical SerDes implementation involves tasks like parallel-to-serial and serial-to-parallel data conversion, impedance matching circuitry incorporation, and clock data recovery functionality integration. The primary objective of SerDes is to minimize the quantity of input/output (I/O) interconnects.

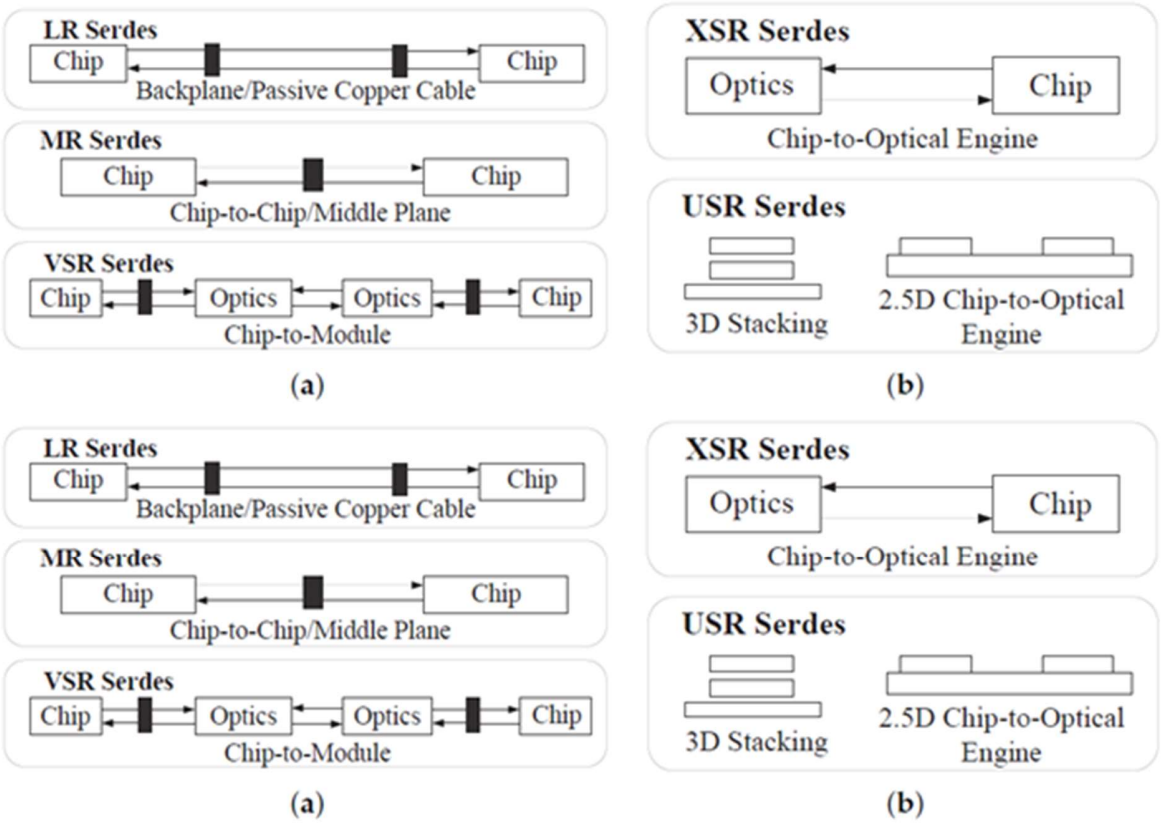


Figure 46 Classification and Application of Typical Serial Interfaces; (a) Classification of Serial Interfaces; (b) Application of Serial Interfaces.

Concerning application transmission distances, SerDes implementations cater to various communication lengths. Serial interfaces encompass Long-Reach (LR), Medium-Reach (MR), and Very Short-Reach (VSR) SerDes. Additionally, Extremely Short Distance (XSR) XSR is for links up to 50 mm, and Ultra-Short Distance (USR) SerDes is designed for extremely close-range communication for links up to 10 mm. XSR and USR SerDes mainly focus on implementing high-speed interconnect communication of Die-to-Die at an ultra-short distance via 2.5D/3D packaging technology. Based on 56 G SerDes interface specification, a comprehensive comparison of multiple interfaces is presented in transmission, application, and other aspects, as shown in Table.

Table 2: Comparisons Among Different Types of OIF 56 G SerDes

Feature	MR/LR SerDes	VSR SerDes	XSR SerDes	USR SerDes
Application Fields	Inter-chip	Chip-to-Module	Die-to-Die & Die-to-Optical Engine	Die-to-Die
Transmission Medium	PCB (1–2 connectors)	PCB (1 connector)	Substrate	Substrate or Silicon Interposer
Coding Scheme	PAM4, ENRZ	PAM4	PAM4, NRZ	NRZ (CNRZ-5)
Bit Error Rate (BER)	1E-4 (1E-10 1E-15 with RS-FEC)	1E-6 (1E-10 1E-15 with RS-FEC)	1E-6	1E-10
Transmission Distance	500–1000 mm	125 mm/25 mm (main PCB/modular PCB)	50 mm	10 mm
Power Consumption per bit	-	-	~5 pJ/bit	~3 pJ/bit

Additional Notes:

- This table compares four different types of OIF 56 G SerDes:
  - MR/LR: Mid-Range/Long-Reach
  - VSR: Very Short Reach
  - XSR: Extremely Short Reach
  - USR: Ultra-Short Reach
- PAM4: Pulse Amplitude Modulation with 4 levels
- ENRZ: Enhanced Non-Return to Zero
- NRZ: Non-Return to Zero
- CNRZ-5: Code-Non-Return to Zero with 5 levels
- RS-FEC: Reed-Solomon Forward Error Correction

The Synopsys XSR PHY intellectual property (IP), designed for 112Gbps per lane die-to-die connectivity, empowers the creation of high-bandwidth interfaces tailored for ultra and extra short-reach applications in multi-chip modules (MCMs). Specifically developed for hyper-scale data centers, artificial intelligence, and networking applications, the XSR PHY IP accommodates NRZ and PAM-4 signaling across data rates ranging from 2.5G to 112 G.

Parallel interface:

Common parallel interfaces for chiplet connections include Intel's AIB/MDIO, TSMC's LIPINCON, and OCP's BoW. The HBM interface serves high-bandwidth storage connections.

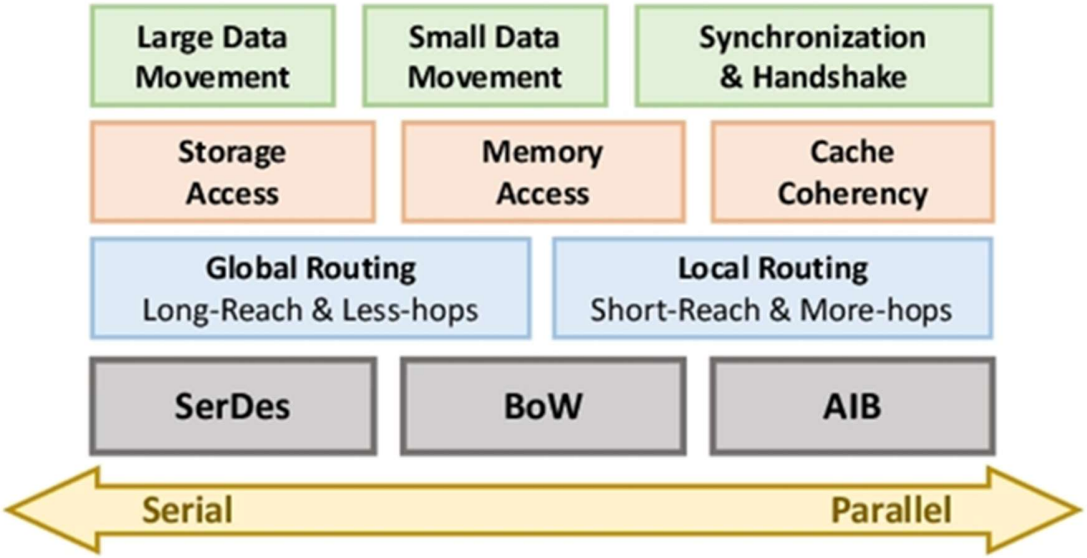
Table 3

Feature	AIB Gen1	MDIO Gen1	LIPINCON	BoW-Turbo (3 slices)
Data Rate (GT/s)	2	5.4	8	16
Shortline Bandwidth Density (Gbps/mm)	504	1600	536	1280
PHY Power (pJ/bit)	0.85	0.5	0.5	0.5 (7 nm)
Package	EMIB	EMIB/ODI	CoWoS	MCP
Areal BW Density (GBps/mm²)	150	198	198	148
Typical Applications	Stratix 10 FPGA	VLSI Presentation	GF Sample	-

Notes:

- GT/s = Gigatransfers per second
- Gbps/mm = Gigabits per millimeter
- pJ/bit = Picojoules per bit
- GBps/mm² = Gigabits per square millimeter
- EMIB = Embedded Multi-die Interconnect Bridge
- ODI = Omni-Directional Interconnect
- CoWoS = Chip-on-Wafer-on-Substrate
- MCP = Multi-chip Package
- This table compares four different parallel interfaces for chiplets: AIB Gen1, MDIO Gen1, LIPINCON, and Bow-Turbo (3 slices).

Intel's AIB is a parallel interconnection standard akin to DDR DRAM Interface. Intel provides free AIB licenses to support the chiplet ecosystem in the CHIPS project. MIDO, an upgraded AIB version, offers over two times the transmission efficiency, speed, and bandwidth. AIB and MDIO suit 2.5D and 3D packaging like EMIB and Foveros. Using advanced packaging tech, TSMC's LIPINCON reduces power and area overhead without PLL/DLL. BoW by OCP ODSA addresses substrate-based interconnection issues with three types: BoW-Base, BoW-Fast, and BoW-Turbo. BoW is versatile, backward-compatible, and has wide applications.



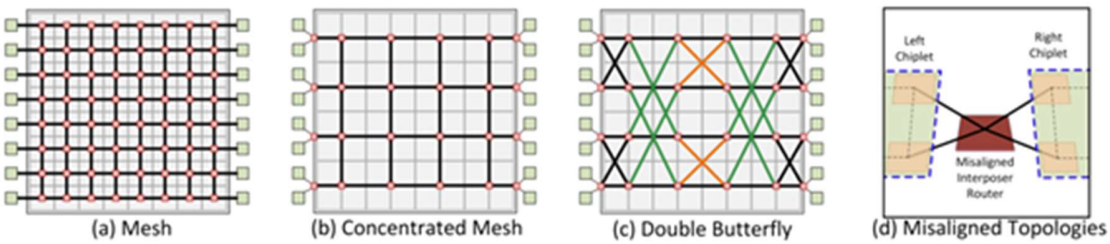
While these interfaces achieve low-power per-bit transmission, growing bandwidth needs pose challenges. Designers must choose interfaces (parallel for low power, low latency, and high bandwidth; serial for fewer resources but more power and delay) based on application requirements, constraints, and die features.

As shown in Figure, different workloads have different requirements for interfaces. For example, the parallel interface is suitable for the frequent local movement of small data, and the serial interface is suitable for the long-distance movement of large data.

Topology for Chiplet architecture:

Terms related to topology:
Hop count: the number of routers or network nodes a data packet passes through to reach its destination. It is a metric used to measure a computer network's distance or routing efficiency, with a lower hop count indicating a more direct and efficient route. Minimizing hop count can help reduce latency and improve overall network performance.
Bisection bandwidth: it measures the total capacity of communication links between two equal halves of a network when it is divided. A higher bisection bandwidth indicates greater network capacity and enhanced parallel communication capabilities.

Interfaces play a vital role in multi-chiplet systems, significantly influencing system design. The chosen interface dictates certain limitations on the system's structure and size. This section explores various topologies. In silicon interposer-based systems, the primary drawback is the expense associated with the interposer and intricate packaging. However, such interfaces may still be employed due to high bandwidth and low latency considerations. These topologies involve communication between different dies within a multi-die system by employing a Network on Interposer (NoI). The chiplets used may have multiple cores and have their own Network on Chip (NoC).

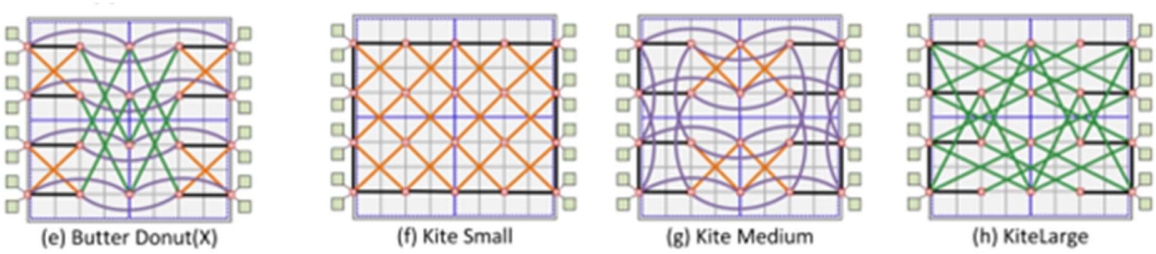


(a) Mesh (b) Concentrated Mesh (c) Double Butterfly (d) Misaligned Topologies

The illustrated design features 64 cores (small grey boxes) organized into 4x4 chiplets (blue dotted lines). Interposer routers, represented by red boxes, connect to the green memory controllers on the side. Different link lengths are employed for each topology, as illustrated by the colored lines (green, orange, purple).
The system builds upon express-link-based topologies and categorizes links into k-straight and k-m-diagonal. In the case of k-straight links, routers are directly linked along the same dimension, bypassing k-1 intermediate routers. For example, 1-straight links denote connections between

nearest neighbor routers, while 2-straight links connect routers, omitting one in the middle (depicted by purple lines). Similarly, k-m-diagonal links adhere to the specified rule, involving k hops in one direction and m in another. Visually, it represents 1-1-diagonal in orange and 2-1-diagonal in green links. The overall operating frequency of the Network-on-Chip (NoI) is influenced by the longest class of links in the topology.

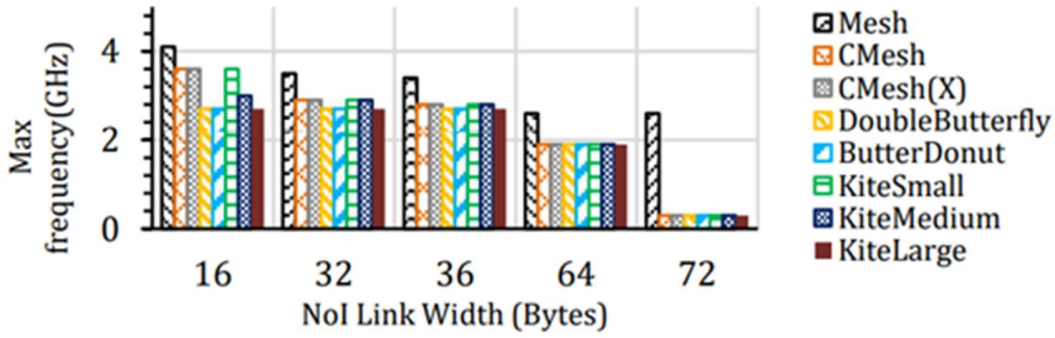
A basic topology links a core-level router to an interposer router, creating a simple mesh structure. However, this configuration takes up a significant area, resulting in the inefficient utilization of numerous links. An alternative solution is the concentrated-mesh network (CMesh), wherein multiple cores connect to a high-radix router in the interposer, reducing the number of network hops. Despite presenting lower throughput compared to a mesh due to a decreased bisection bandwidth, CMesh has its own advantages. The Double Butterfly topology employs longer links to minimize the average hop count, with diagonal links enhancing the bisection bandwidth.


(e) Butter Donut(X) (f) Kite Small (g) Kite Medium (h) KiteLarge

To optimize the network, "misaligned" can be introduced where router locations are deliberately offset to enable cores on the periphery of adjacent chiplets to share the same router. This allows simultaneous chip-to-chip and memory traffic flow through a router, reducing queuing delays for messages traversing the network's bisection cut. The ButterDonut topology refines the misaligned approach to minimize the average hop count and latency in inter-chiplet coherency traffic. Depending on the chosen topology, interconnect misalignment can be implemented in the X, Y, or both dimensions (in the plane parallel to the interposer). ButterDonut enhances bisection bandwidth while maintaining a router complexity comparable to CMesh, and ButterDonut(X) specifically denotes misalignment in the X dimension.

Within an interposer-based system, chiplets and the interposer can possess individual interconnect networks, functioning at distinct clock domains and/or link widths. This results in a heterogeneous interconnection fabric comprising Network-on-Chip (NoC) and Network-on-Interposer (NoI) across the system. The combination of NoC and NoI in these hybrid architectures presents new design challenges, encompassing routing, composability, yield, and thermal considerations. The Kite topology can be employed when NoC and NoI are decoupled. Kite topologies utilize the misaligned NoI configuration to minimize the hop count for inter-chiplet traffic. To illustrate, three different link lengths (1-1-diagonal, 2-straight, and 2-1-diagonal) are employed, resulting in three distinct topologies: Kite-Small, Kite-Medium, and Kite-Large, each utilizing these link lengths as the longest links, respectively.

Technology-Specific Evaluation

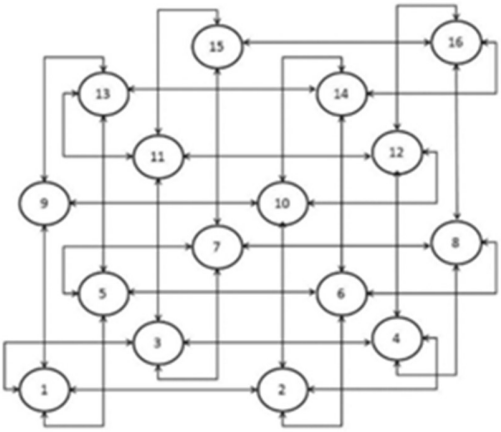


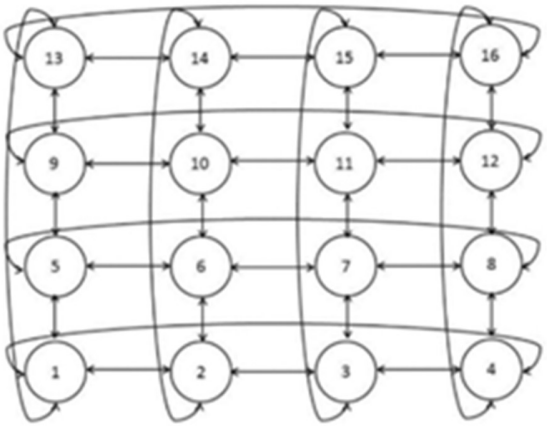
The above graph shows the Maximum Operating Frequency for NoI topologies with various link widths at the 22nm technology. The maximum operating frequency is determined by the radix of the largest router and the length of the longest wire. The mesh topology, characterized by shorter links and low radix routers, can achieve an operational frequency of 4.0 GHz. CMesh exhibits resilience up to 3.6 GHz. However, topologies with longer links, such as Double Butterfly and ButterDonut, are constrained to a maximum frequency of 2.7 GHz. The figure also illustrates that an increase in link width results in a reduction of the maximum operating frequency.

The table summarizes NoI topologies, covering a range of metrics, both specific to the topology and the technology employed. The average hop count, serving as a design-time measure for latency independent of runtime traffic contention, is calculated by averaging hops across all possible source-destination pairs ($Havg$). The findings indicate that, despite certain topologies like CMesh having a high average hop count, they might outperform others when considering their maximum operating frequencies. Conversely, while alternative topologies boast a lower average hop count, they encounter a drawback due to their longer links' need to operate at a lower clock frequency.

Topology	Topology Specific						Technology Specific	
	Routers	Links	Diameter	Avg Hop Count H_{avg}	Bisection BW	Router Deg.	Longest Link (mm)	Max Freq (GHz)
Mesh	64	112	16	6.12	8	5	2.2	**4.0**
CMesh	24	32	8	3.75	4	8	4.4	**3.6**
CMesh(X)	**20**	40	7	3.25	4	8	4.4	**3.6**
Double Butterfly	24	40	4	2.85	8	8	9.8	**2.7**
Butter Donut	**20**	36	4	2.21	**12**	8	9.8	**2.7**
Kite Small	**20**	38	4	2.39	8	8	6.2	**3.6**
Kite Medium	**20**	40	4	**2.17**	**12**	8	8.8	**3.0**
Kite Large	**20**	36	**3**	**2.03**	**12**	8	9.8	**2.7**

Torus and Fold torus topology





The Torus topology, characterized by long wrap-around links connecting boundary nodes in the same row and column, is popular for its advantages in reduced network diameter and hop count. However, it requires long wires for wrap-around links compared to a mesh. The Folded Torus, as shown in the figure, improves on this by offering shorter link lengths, minimizing packet traversal time, and reducing the necessary interconnect area. Additionally, the Folded Torus provides enhanced path diversity and fault tolerance compared to the Torus. Misalignment can be added to improve its performance.

Table 4: Comparisons of Chiplet Packaging Technologies

Feature	Substrate-Based	Silicon Interposer-Based	RDL-Based	Fan-Out
Integration Density	Low	High	Middle	High
Transmission Performance	Low	High	High	High
Routing Resources	Highest	Higher	Middle	High
Heat Dispersion	Middle	Middle	High	High
Cost	Low	High	Middle	Low
3D Extensibility	Low	Middle	High	High
Provider	Chip Packaging Test Factory	Chip Foundry	Packaging Test Factory/Foundry /Integrated Device Manufacture	Foundry

Notes:

- Substrate-based: This technology uses a simple organic or silicon substrate to connect the chiplets. It is inexpensive but has limitations in density, routing resources, and thermal performance.
- Silicon interposer-based: This technology uses a silicon interposer as a high-performance interconnection layer. It offers higher density, routing resources, and thermal performance than substrate-based packaging. However, it is more expensive.
- RDL-based: This technology uses a redistribution layer (RDL) on top of a silicon substrate to connect the chiplets. It offers a good balance of performance and cost.
- Fan-out: This technology uses a special silicon carrier to connect the chiplets. It offers high density and routing resources but can be expensive.
- Integration density: This refers to the number of chiplets that can be integrated into a package.
- Transmission performance: This refers to the speed and bandwidth of the connections between the chiplets.
- Routing resources: This refers to the number of available routing channels for connecting the chiplets.
- Heat dispersion: This refers to the ability of the package to dissipate heat generated by the chiplets.
- Cost: This refers to the relative cost of the packaging technology.
- 3D extensibility: This refers to the ability to stack multiple chiplets in a 3D configuration.

Characterizing and Analyzing Die-To-Die Channels in 2.5D and 3D MCM Architectures

HSPICE simulations were used to assess die-to-die communication channels in 2.5-D and 3-D heterogeneous integration platforms. The study includes the evaluation of delay, energy-per-bit, and bandwidth-density for these integration platforms. Heterogeneous Interconnect Stitching Technology (HIST) in 2.5-D integration demonstrates a maximum latency and energy reduction of 6.2% and 15.1%, respectively, compared to other 2.5-D integrated systems with a 1 mm interconnect length. Furthermore, 3-D ICs exhibit enhanced performance, with link latency and energy approximately 19.4% and 48.0% smaller than those of HIST (1 mm wire) for TSV-based 3-D integration (75 µm TSV height).

In order to simulate circuits at such a small level we need to take parasitic capacitances into account, now how do we do that?

We referenced papers and gathered these Quantifications for parasitic capacitances.

MODELS FOR PARASITIC ESTIMATION

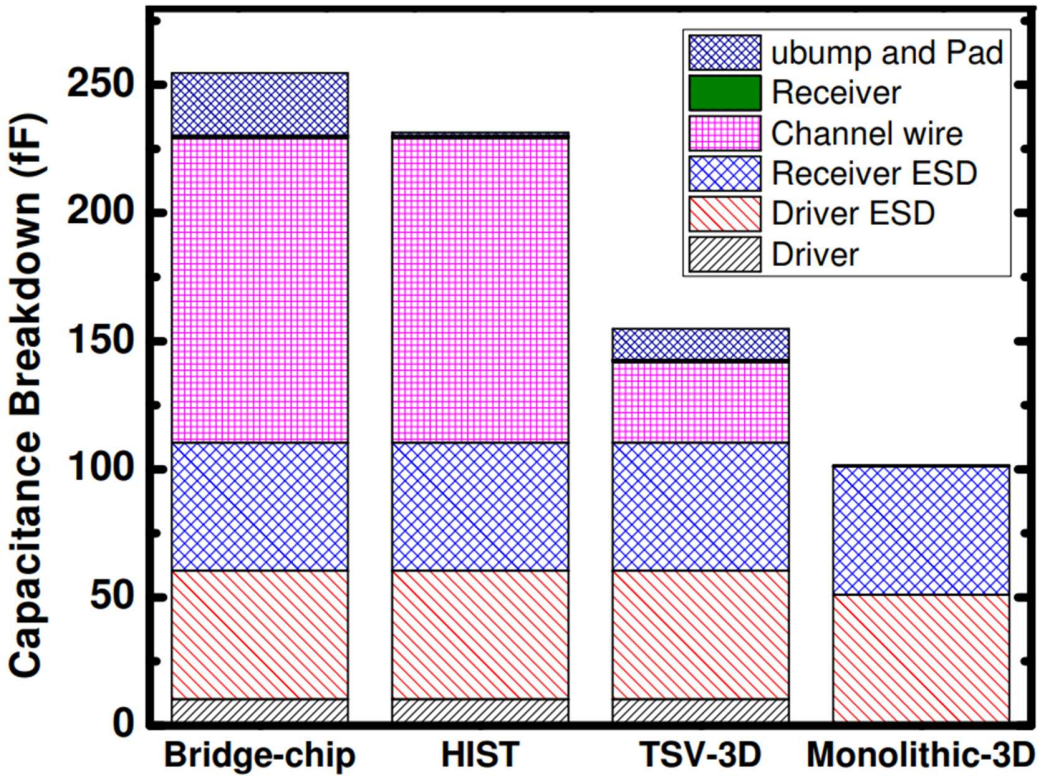
Component	Equation
Wire to substrate capacitance: C_{wire} [12]	$C_{wire} = \epsilon_0 \cdot \epsilon_{ox} \cdot \frac{W \times L}{t_{ox}}$ W and L are wire width and length.
Wire to wire coupling capacitance: C_{wire} [12]	$C_{wire} = \epsilon_0 \cdot \epsilon_{ox} \cdot \frac{t_{ox} \times L}{P_{wire}}$ P_{wire} is the wire pitch
Pad capacitance: C_{pad} [14]	$C_{pad} = \epsilon_0 \cdot \epsilon_{ox} \cdot \frac{W_p^2}{t_{ox}}$ W_P is the pad width and t_{ox} is the ILD thickness
Microbump to ground capacitance: C_{bump} [14], [16]	$C_{bump} = \epsilon_0 \cdot \epsilon_{underfill} \cdot \frac{2 \cdot \pi \cdot H_{bump}}{arcosh(\frac{P_{bump}}{D_{bump}})}$ P_{bump} is the microbump pitch
Microbump inductance: L_{bump} [14]	$L_{bump} = \frac{\mu_0 \cdot \mu_{bump}}{2 \cdot \pi} \times H_{bump} \times ln(\frac{2 \cdot P_{bump}}{D_{bump}})$
TSV dioxide capacitance: C_{ox} [14]	$C_{ox} = \epsilon_0 \cdot \epsilon_{ox} \cdot \frac{2 \cdot \pi \cdot H_{tsv}}{ln(\frac{D_{tsv}}{D_{tsv}-2 \cdot t_{ox}})}$ t_{ox} is the dioxide thickness, H_{tsv} is the TSV height
TSV depletion capacitance: C_{dep} [16]	$C_{ox} = \epsilon_0 \cdot \epsilon_{Si} \cdot \frac{2 \cdot \pi \cdot H_{tsv}}{ln(\frac{D_{tsv}+W_{dep}}{D_{tsv}}}$ W_{dep} is the depletion width
TSV total capacitance: C_{tsv}	$C_{tsv} = \frac{C_{ox} \cdot C_{dep}}{C_{ox} + C_{dep}}$
TSV inductance: L_{tsv} [14]	$L_{tsv} = \frac{\mu_0 \cdot \mu_{tsv}}{2 \cdot \pi} \times H_{tsv} \times ln(\frac{2 \cdot P_{tsv}}{D_{tsv}})$
TSV to substrate capacitance: C_{Si} [16]	$C_{ox} = \epsilon_0 \cdot \epsilon_{Si} \cdot \frac{2 \cdot \pi \cdot H_{tsv}}{arcosh(\frac{P_{tsv}}{D_{tsv}})}$ t_{ox} is the dioxide thickness, H_{tsv} is the TSV height
TSV to substrate resistance: R_{Si} [16]	$R_{Si} = \frac{\epsilon_0 \cdot \epsilon_{si}}{C_{Si} \cdot \sigma_{Si}}$ σ_{Si} is the silicon conductance

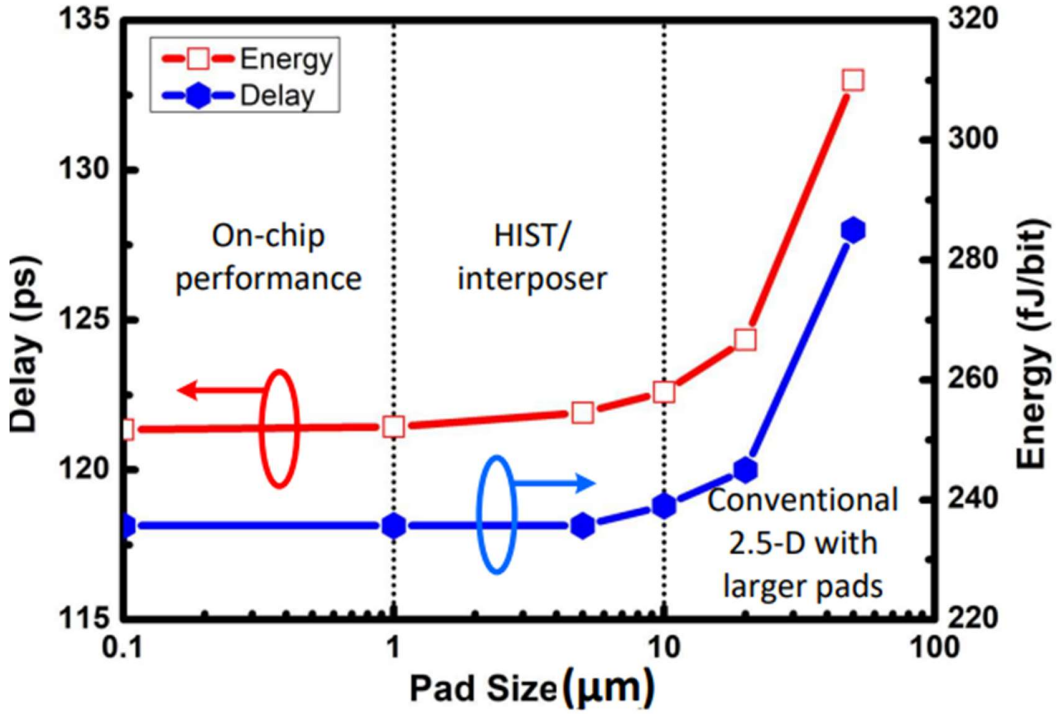
PHYSICAL DIMENSIONS CONSIDERED IN THE SIGNALING MODELS

Parameter	value
Link wire length (mm)	$0.1 \sim 5$
Link wire pitch/thickness/width (μm)	1.6/2/0.8
Pad width (μm)	$0.1 \sim 50$
Microbump diameter & height	$0.8 \times$ Pad width
TSV diameter (μm)	$0.1 \sim 10$
TSV height	$15 \times$ TSV diameter
TSV dioxide thickness	$0.05 \times$ TSV diameter
Microbump/TSV pitch	$2 \times$ Microbump diameter

Here are some of the terms explained:

- **Wire to Substrate Capacitance (Cwire):**
  - This represents the capacitance between a wire and the substrate (the underlying material). It depends on the characteristics of the wire, such as its width and length.
- **Wire to Wire Coupling Capacitance (Cwire):**
  - This refers to the capacitance between two adjacent wires. It is influenced by the characteristics of the wires, such as their length and pitch (the distance between corresponding points on adjacent wires).
- **Pad Capacitance (Cpad):**
  - This is the capacitance associated with the electrical pads. Pads are components used for connecting external devices or components to the integrated circuit. The capacitance depends on the pad width and the interlayer dielectric (ILD) thickness.
- **Microbump to Ground Capacitance (Cbump):**
  - This is the capacitance between a microbump (small metallic connection) and the ground. Microbumps are used for connecting different layers in integrated circuits. The capacitance depends on the microbump pitch and the material properties.
- **Microbump Inductance (Lbump):**
  - This is the inductance associated with a microbump. Inductance is a measure of how much a component resists changes in the electric current flowing through it. It depends on the microbump's dimensions and material properties.
- **TSV Dioxide Capacitance (Cox):**
  - This represents the capacitance associated with the Through-Silicon Via (TSV) and its oxide layer. TSVs are vertical electrical connections passing through a silicon wafer. The capacitance depends on the oxide thickness and the dimensions of the TSV.


Capacitance Breakdown (fF)
250
200
150
100
50
0
ubump and Pad
Receiver
Channel wire
Receiver ESD
Driver ESD
Driver
Bridge-chip
HIST
TSV-3D
Monolithic-3D


Energy
Delay
On-chip performance
HIST/ interposer
Conventional 2.5-D with larger pads
Delay (ps)
Energy (fJ/bit)
Pad Size(μm)
135
130
125
120
115
320
300
280
260
240
220
0.1
1
10
100

Feature	Bridge-chip	Interposer	HIST	TSV-3D	Monolithic-3D
Microbump (pitch/diameter/height)	50/25/50 µm†	(8 ~ 24)//(4 ~ 12)/8/4/4 µm	40/20/20 µm	0.4/0.2/0.2 µm	(4 ~ 12) µm
Pad size	37.5 µm	5 ~ 15 µm	5 µm	30 µm	0.3 µm
Link wire length	1 mm	1 mm	1 mm	75 µm	800 nm
Microbump capacitance	14.75 fF	1.18 ~ 3.54 fF	1.18 fF	5.90 fF	0.06 fF
Pad capacitance	9.72 fF	0.17 ~ 1.56 fF	0.17 fF	6.20 fF	~ 0 fF
Link wire capacitance	118.9 fF	118.9 fF	118.9 fF	31.5 fF	0.1 fF
Link latency with ESD	125.1 ps	117.3 ~ 118.6 ps	117.3 ps	98.9 ps	94.5 ps
Link energy with ESD	306.3 fJ/bit	259.9 ~ 267.4 fJ/bit	259.9 fJ/bit	176.2 fJ/bit	135.1 fJ/bit
Link latency without ESD	107.9 ps	99.7 ~ 101.0 ps	99.7 ps	75.9 ps	33.0 ps
Link energy without ESD	206.0 fJ/bit	159.5 ~ 167.0 fJ/bit	159.5 fJ/bit	76.2 fJ/bit	3.7 fJ/bit

Simulation and Results:

- Circuit models are simulated using HSPICE, and latency and energy results are obtained for all considered heterogeneous integration scenarios.
- HIST and interposer demonstrate better electrical performance than the bridge-chip approach due to smaller pads and microbumps.
- The total capacitance of a microbump and a pair of pads for HIST is significantly smaller than that of bridge-chip, resulting in reduced latency and link energy.
- HIST achieves a 6.2% reduction in latency and approximately 15.1% reduction in link energy compared to bridge-chip.
- The impact of pad size on channel latency and energy dissipation is studied, revealing diminishing returns below a certain pad size.
- 3-D integration, with significantly shorter die-to-die wires, shows smaller link latency and energy than 2.5-D designs.
- TSV-based 3-D design achieves approximately 15.7% smaller latency and 32.2% smaller energy than HIST/interposer.
    - Monolithic 3-D ICs, using nanoscale vertical vias, show even more substantial reductions in latency and energy compared to HIST/interposer.

"HIST" stands for "Heterogeneous Interconnect Stitching Technology." HIST is an approach or technology used in the field of integrated circuits for heterogeneous integration.

However, for previously mentioned reasons in the Mid Submission Reports other architectures than 2.5D were less feasible for JLR, since they sacrificed swappability, ease of production, and thermal management.

Samuel Naffziger

AMD SVP

These small die were a huge enabler for us, I view this as one of the greatest engineering achievements in the industry and in recent memory because it solves so many problems at once.

THERMAL MANAGEMENT

A THERMALLY AWARE CHIPLET PLACEMENT FOR 2.5D SYSTEM

Chiplets are densely packed in the conventional layout of a 2.5D heterogeneous system to reduce wirelength, however this setup tends gets poor heat management.

We present TAP 2.5D(Thermally Aware Placement): This methodology focused on placement of chiplet that takes into account thermal considerations, strategically introducing space between chiplets to lower temperatures and minimize overall wirelength.

TAP-2.5D is designed specifically for 2.5D heterogeneous integration with a particular connectivity in network. Its goal is to optimize the arrangement of chiplets while simultaneously reducing operational temperatures and the total length of wires used.

Placement Description:

To illustrate unrestricted placements, we consider the chiplets' widths, heights, and their central x and y coordinates. We have the interposer segmented into a discrete grid “Occupation Chiplet Matrix (OCM)”. Our approach confines a chiplet's center to the grid's intersection nodes, preventing an infinitely expansive solution space.

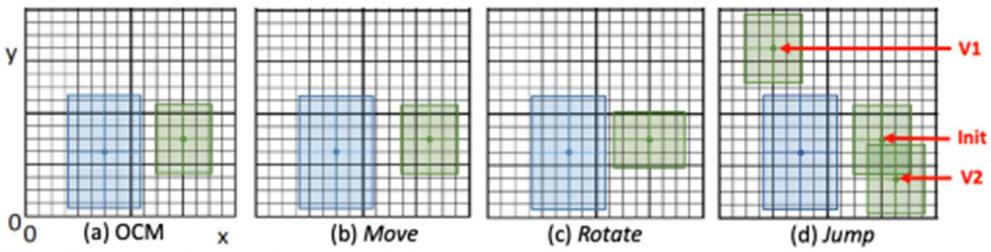


Figure 47 (a) A logical view of the OCM data structure used by TAP-2.5D modeling two chiplets over the floorplan. (b, c, d) represent three examples of chiplet movements – V1 and V2 are two valid positions for the jump operation starting from the initial chip let placement (Init).

We start with a compact chiplet placement solution derived from the B*-tree and fast-SA based method as the initial layout.

To generate a neighbour placement, we modify the current chiplet arrangement using move (Fig.(b)), rotate (Fig.(c)), and jump (Fig.(d)) operations, ensuring the result is a valid placement.

- For rotation, we select a chiplet randomly and rotate it by 90 degrees.
- In the case of a move, we randomly choose a chiplet and shift it minimally (1 mm in our scenario) up, down, left, or right, ensuring no overlaps occur post-movement.

- The likelihood of chiplet positions changing significantly using only rotate and move operations is uncertain. This could lead to the "sliding tile puzzle" problem, where chiplets might get stuck due to others blocking their movement, occurring during the SA process.
- To address this 'sliding tile puzzle' concern, we employ the jump operation. This allows a randomly chosen chiplet to leap to any valid empty spot on the interposer. A valid neighboring placement must ensure chiplets don't overlap and stay entirely within the interposer boundaries.

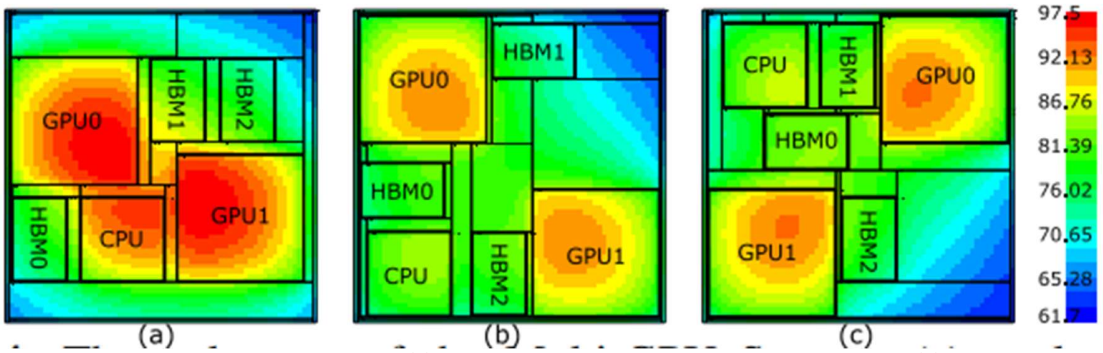


Figure 48 : Thermal maps of the Multi-GPU System: (a) a placement solution using Compact-2.5D approach, (b) TAP-2.5D solutions using repeaterless non-pipelined inter-chiplet links, and (c) gas-station links.

- The placement shown in Fig.(a) is achieved through the Compact-2.5D method, prioritizing wirelength and area minimization without considering temperature. This configuration functions at 95.31oC, with a total wirelength (the sum of all inter-chiplet link lengths) measuring 88,059mm.
- The configuration depicted in Fig.(b) is the result of our TAP-2.5D technique, employing a physical network featuring repeaterless non-pipelined inter-chiplet links. This arrangement exhibits a reduced peak temperature of 91.25oC; however, it comes with an extended total wirelength of 96,906mm due to the positioning of high-power CPU and GPU chiplets in the corners.
- Our placement solution with the gas-station links is shown in Fig. (c). The system's temperature is similarly lower (91.52oC) but the total wirelength reduces to 51, 010mm (vs. 88, 059mm obtained by Compact-2.5D).

TAP 2.5D deliberately adds space between chiplets to enhance heat dissipation and raises the system's total thermal design power. Our method involves simulating annealing and looks for a chiplet placement that considers temperature. It also optimizes the routing of inter-chiplet wires in heterogeneous 2.5D systems.

One another solution could be bumpless hybrid bonding.

- Bumpless HB, also known as low-temperature direct bond interconnect, involves a permanent bonding process.

- At typical temperatures, the dielectric-to-dielectric (often SiO2-SiO2) naturally forms a bond. This necessitates dielectric surfaces with extremely minimal roughness. Subsequently, a low-temperature batch annealing method is utilized to join the metal-to-metal (usually Cu-Cu) on the opposing side of the wafer or die.
- HB technology eliminates the need for bumping and underfilling, thus avoids associated thermal and thermomechanical issues. It also enables finer pitch, greater bandwidth, improved thermal performance, reduced parasitic parameters, and increased assembly throughput. Consequently, it stands as the most promising vertical interconnect technology for chipset packaging in the upcoming years.

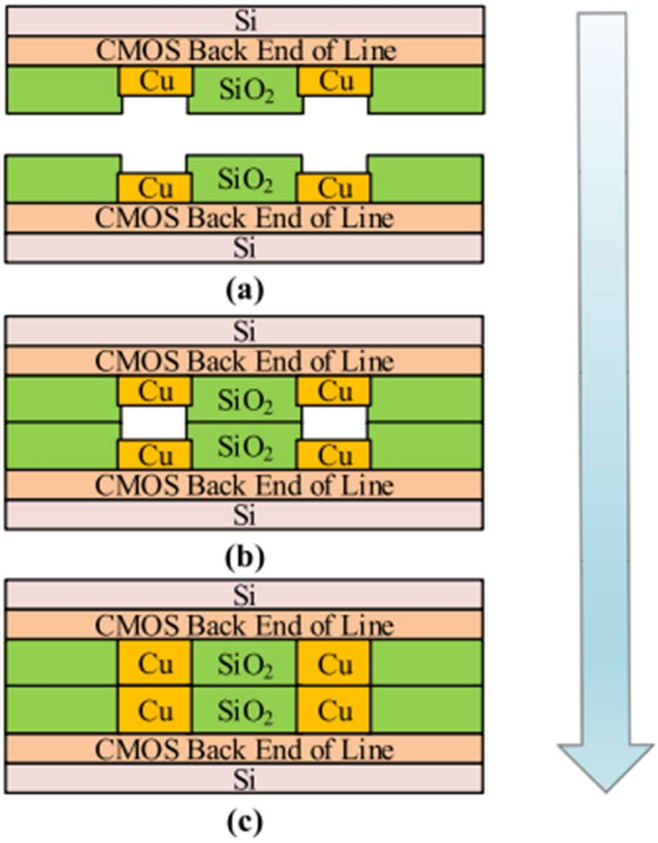


Figure 49 Critical process steps of Chip to Wafer/Wafer to Wafer HB. (a) Metal access and plasma surface activation. (b) SiO2-to-SiO2 initial bond at room temperature. (c) Cu–Cu bond at low-temperature annealing.

It primarily involves three key stages.

Stage	Figure	Description	Key Actions	Potential Issues
Preparation (Fig. a)	(a)	Sets foundation for bonding	- CMP polishes surfaces. - Cu dishing occurs (faster erosion than SiO2). - Surface roughness and flatness ensured.	- Excessive Cu dishing could weaken bond. - Roughness/unevenness affect bond strength.

Direct Bonding (Fig. b)	(b)	Interlayer dielectric fusion	- Bonding at ambient temperature. - Top and bottom surfaces fuse.	- Insufficient fusion leads to weak bond. - Temperature control crucial.
Post-Bonding Annealing (Fig. c)	(c)	Closes Cu dishing gap	- Controlled temperature ramp-up and dwell. - Cu pad protrusion fills the gap.	- Excessive temperature damages device. - Incomplete gap closure compromises bond.

Given Cu's considerably higher coefficient of thermal expansion compared to SiO2, Cu expands significantly, bridging the spaces between two metal surfaces and resulting in a permanent bond between Cu–Cu surfaces.

COOLING TECHNIQUES

Due to the extensive integration in the Chiplet Heterogeneous Integration (CHI) system, significant power consumption occurs, converting into heat. Additionally, the limited design space in chiplet arrays aims to minimize wire length, intensifying the challenge of heat dissipation. Failure to address thermal issues adequately can result in many drawbacks caused by overheating.

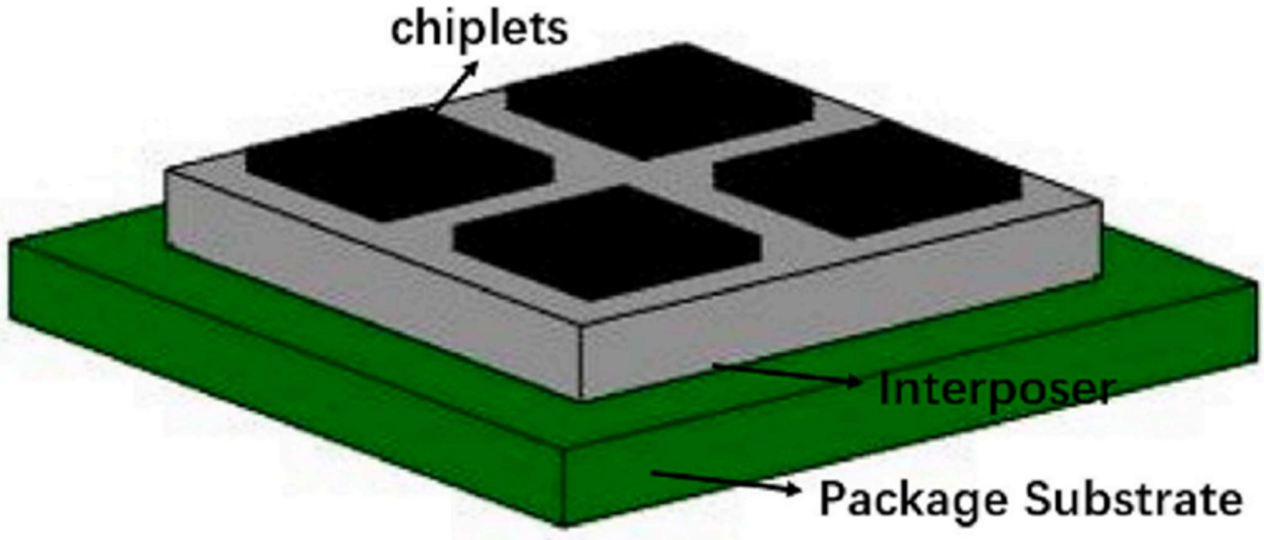


Figure 50 Framework of CHI(Chiplet Heterogeneous Integration)

Factors for thermal consideration:

- Increased expenses associated with the development of monolithic silicon.
- The intricate manufacturing process involved in creating 3D stacked dies.
- Concerns regarding hot spots or the maximum allowable operating temperatures of ASIC and SRAM within MCM or 2.5D modules.

Microchannels:

- Due to frequent conflicts among new technologies, creating chips that accommodate diverse functions (like memory, arithmetic units, and sensors) on the same substrate is unfeasible. Consequently, modern systems comprise an increasing number of components packaged together in current 2.5-D or 3-D electronics packaging.

- Conventional methods face challenges in efficiently transferring heat from junctions to the ambient environment. This heat transmission occurs through semiconductor dies and the lid of a flip chip ball grid array (FCBGA) package, among other pathways.

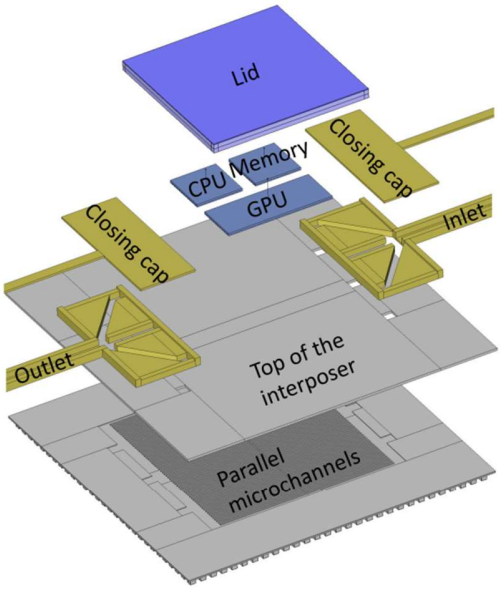


Figure 51 Exploded view of heterogeneous packing

- In the realm of modern System-on-Chip (SoC) devices, flip-chip packaging technology is commonly employed. The primary heat path extends from the junction to the chip's backside and further to the heatsink and fan (HSF) assembly through the semiconductor die.

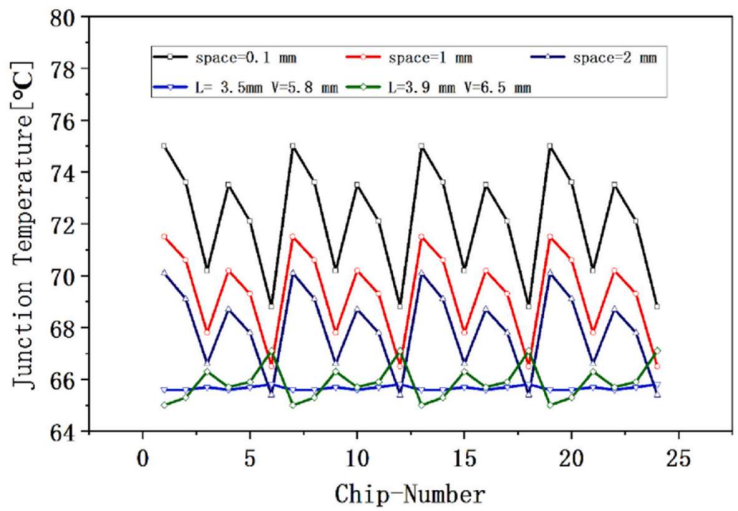


Figure 52 Junction temperature of each chiplet under different space parameters

We have statistically evaluated a unique microgap dielectric coolant manifold for 2.5-D stacked integrated circuits with numerous high-power dies.

- A monolithic microfluidic cooling system designed for 2.5-D packages employed deionized water as a coolant within micropin fin heatsinks carved directly onto the backside of the chiplets and shielded by a 3-D printed manifold.

- To ensure minimal junction-to-ambient thermal resistance, we developed a 3-D-printed cold plate. This was placed atop a device (a copper block simulating a computer chip, measuring 1" × 1" in top surface area) and characterized its performance using water as the coolant.

One possible way is to enhance the role or create a secondary heat flow path by applying microchannel-based cooling.

To improve the transfer of heat from the junction to the board, we aim to optimize by incorporating microchannel structures into the interposer layer within a straightforward stack-up.

These microstructures are created directly on the backside of the active silicon semiconductor dies using CMOS-compatible wet chemical etching processes.

Additionally, there exists a secondary heat flow pathway between the active region and the environment. This pathway involves heat traveling through bumps, the interposer, microbumps, and eventually reaching the mainboard, which functions akin to a heatsink in this scenario.

This approach allows us to reduce the number of thermal interface material (TIM) layers, minimizing the heat conduction path between the junction and the microchannel wall where heat exchange occurs between the die and the circulating fluid material.

In scenarios where the die is placed on a lead frame with electrical connections established using different wirebond techniques, heat can solely dissipate to the surroundings via natural convection and conduction (moving through the lead frame towards the pins). The thermal resistance R_{thja} can be notably decreased by introducing an additional heat flow route, especially if microchannels are present.

In scenarios where the die is placed on a lead frame with electrical connections established using different wirebond techniques, heat can solely dissipate to the surroundings via natural convection and conduction (moving through the lead frame towards the pins). The thermal resistance R_{thja} can be notably decreased by introducing an additional heat flow route, especially if microchannels are present.

Flip-chip and wirebond procedures are often employed on a common interposer in the case of 2.5-D or 3-D packaging. Nevertheless, it is not possible to use the conventional FCBGA package structure. The height differences between the dies on the same interposer are the main ones. In this instance, applying TIM layers of rent thicknesses is necessary to achieve thermal contact between the die and the common metal lid/IHS.

For 2.5-D packaging and heterogeneous integration, an interposer—either silicon or organic—can accommodate multiple dies with varying thicknesses and bonded using different techniques like flip-chip or wire bond. Placing microchannels close to the junction is crucial to minimize the overall thermal resistance between the junction and the ambient. The method previously discussed for creating embedded microchannel structures in a silicon die is also suitable for silicon interposers.

Typically ranging from 20 to 30 mm, the interposer's side edges are significantly larger than those of the chips, approximately ten times larger. Consequently, the channel lengths are also extended compared to when channels are realized within the chip. This extension leads to higher pressure drops and increased hydrodynamic resistance, necessitating the use of additional or larger and deeper channels.

It is also important to remember that, in the case of parallel channels, each channel's dimensions—that is, its length and characteristic dimensions—should match those of the others.

It was required to divide each channel into manageable chunks and determine the Nusselt number and heat transfer coefficient values for each chunk to derive the analytical model of heat transfer between the channel walls and the coolant.

Near the inlets, the local Nusselt number and, consequently, the heat transfer coefficient, exhibit significantly higher values. The formula for the heat transfer coefficient, denoted as 'h,' is expressed as follows:

$$h = \frac{k_f \cdot Nu}{D_H}$$

Figure 53 Heat Transfer Coefficient

Where: k_f is the thermal conductivity of the fluid, and D_H is the hydraulic diameter.

D_H hydraulic diameter can be determined for a square-based column as follows:

$$D_H = \frac{2 \cdot a \cdot b}{a + b}.$$

Figure 54 Hydraulic Diameter

The Nusselt number can be calculated as follows:

$$\mathrm{Nu} = 5.14 + \frac{0.065 \cdot (D_H/L) \cdot \mathrm{Re} \cdot \mathrm{Pr}}{1 + 0.04 \cdot \left[(D_H/L) \cdot \mathrm{Re} \cdot \mathrm{Pr}\right]^{\frac{2}{3}}}$$

Figure 55 Nusselt Number

Where, L is the length of the channel, R_e is the Reynolds number, and P_r is the Prandtl number, which can be calculated as follows:

$$\mathrm{Re} = \frac{D_H \cdot \frac{\mathrm{d}m}{\mathrm{d}t}/A}{\mu} = \frac{D_H \cdot \rho \cdot \frac{\mathrm{d}V}{\mathrm{d}t}/A}{\nu} = \frac{D_H \cdot \rho \cdot v}{\nu}$$

$$\mathrm{Pr} = \frac{c_p \cdot \nu}{k_f}$$

Figure 56 Reynolds Number and Prandtl Number

Where, ν is the dynamic, μ is the kinematic viscosity of the fluid, ρ is the density of the fluid, L is the length of the investigated channel, and v is the mean velocity of the fluid.

When considering only a segment of the channel measuring dx in length, heat travels through the channel walls, heating the fluid. Using the principle of energy conservation, we can express this relationship.

$$h \cdot \left(T_w - T_f\right) \cdot p \cdot \mathrm{d}x = \frac{\mathrm{d}m}{\mathrm{d}t} \cdot c_p \cdot \mathrm{d}T_f$$

Figure 57 Conservation of Energy Equation

Where, c_p is the specific heat of the fluid, dT_f is the temperature difference of the dm mass fluid when it enters and leaves the dx length segment of the investigated channel per unit time, dm/dt is the constant mass flow rate of the fluid, and T_w and T_f are the walls and fluid temperature, respectively.

By resolving the resultant linear differential equation, we can compute the thermal resistance between the walls of an individual channel and the fluid.

$$R_{\mathrm{th\_uch}} = \frac{1}{\frac{\mathrm{d}m}{\mathrm{d}t} \cdot c_p \cdot \left(1 - e^{-\frac{h \cdot A}{\frac{\mathrm{d}m}{\mathrm{d}t} \cdot c_p}}\right)}.$$

Figure 58 Thermal Resistance

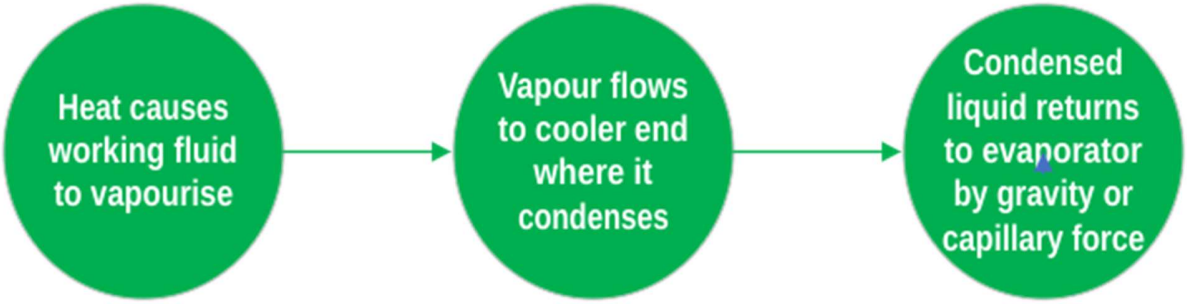
The total $R_{th\_uch\_str}$ value for parallel microchannel structures can be found by summing the G_{th_uch} conduction values of each channel and then computing its reciprocal.

$$R_{\mathrm{th\_uch\_str}} = \frac{1}{\sum_{i=1}^{N} G_{\mathrm{th\_uch}\_i}}.$$

Figure 59 Thermal Resistance

HEAT PIPES

- It is a passive two-phase heat transfer divide operating in a closed system.
- A heat pipe efficiently handles thermal management by utilizing a working fluid within a vacuum.
- Heat is transferred through an evaporator, causing the fluid to boil and turning it into vapor, which then moves to the cooler area of the heat pipe. The colder region is connected to a heat sink where the vapors release their heat and condense back into a liquid.
- This process minimizes the temperature difference within the heat pipe significantly.



- Heat pipes are light weight and can reduce overall system weight.
- They are flexible and can fit countless geometries.

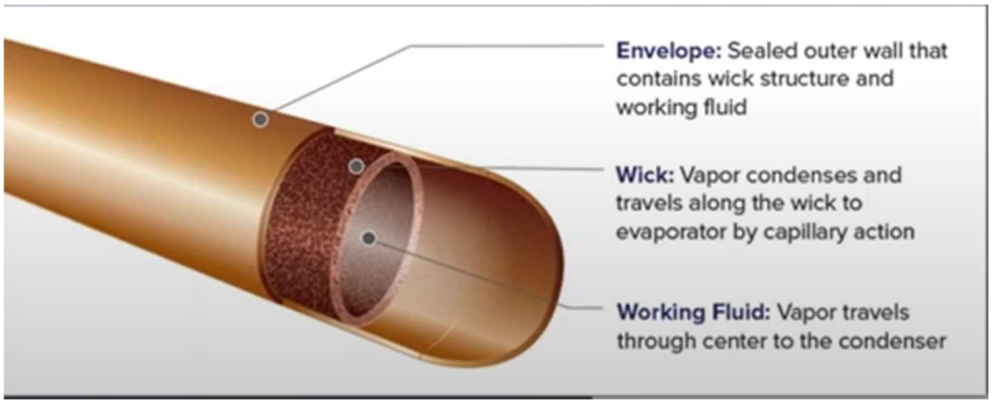


Figure 60 Heat Pipe Structure

Some of the advantages of using heat pipe are:

1. Requires no mechanical or electrical input.
2. Does not involve any moving part.
3. Maintenance free.

Heat pipes can withstand harsh environments like shock and vibration hence can be used in automobiles.

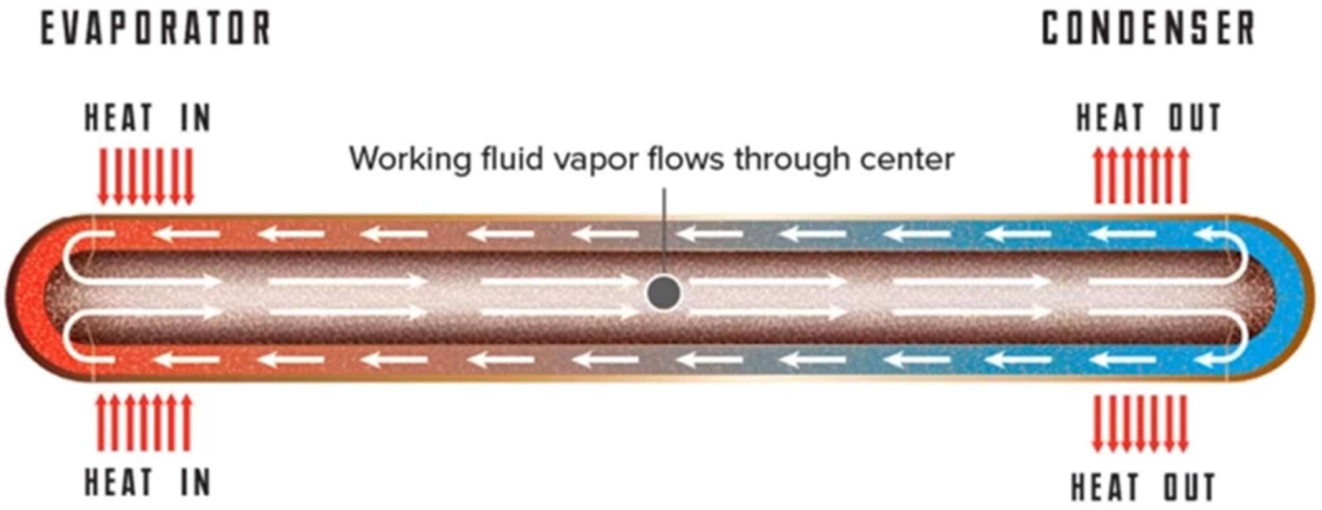


Figure 61 Working of Heat Pipe

VAPOUR CHAMBER

- Compared to a solid conduction heat spreader, a vapour chamber has lower resistance and more consistent temperature, making it a preferred heat spreader for high heat flux cooling applications.
- Heat is delivered to the liquid at the bottom of a vapor chamber, where it evaporates, travels a short distance, and condenses on the chamber's top plate.
- A Vapour Chamber includes an envelope, a wick structure, and a working fluid and uses evaporation, condensation, and capillary transport of the working fluid to provide a high effective thermal conductivity.
- A Vapour Chamber can be described as a closed hollow object that is filled with working fluid, generally at a vacuum pressure (below atmospheric).
- Vacuum pressure serves the aim of reducing the boiling point of the chosen refrigerant, enabling operation at an intermediate temperature between the allowable evaporator and condenser temperatures.

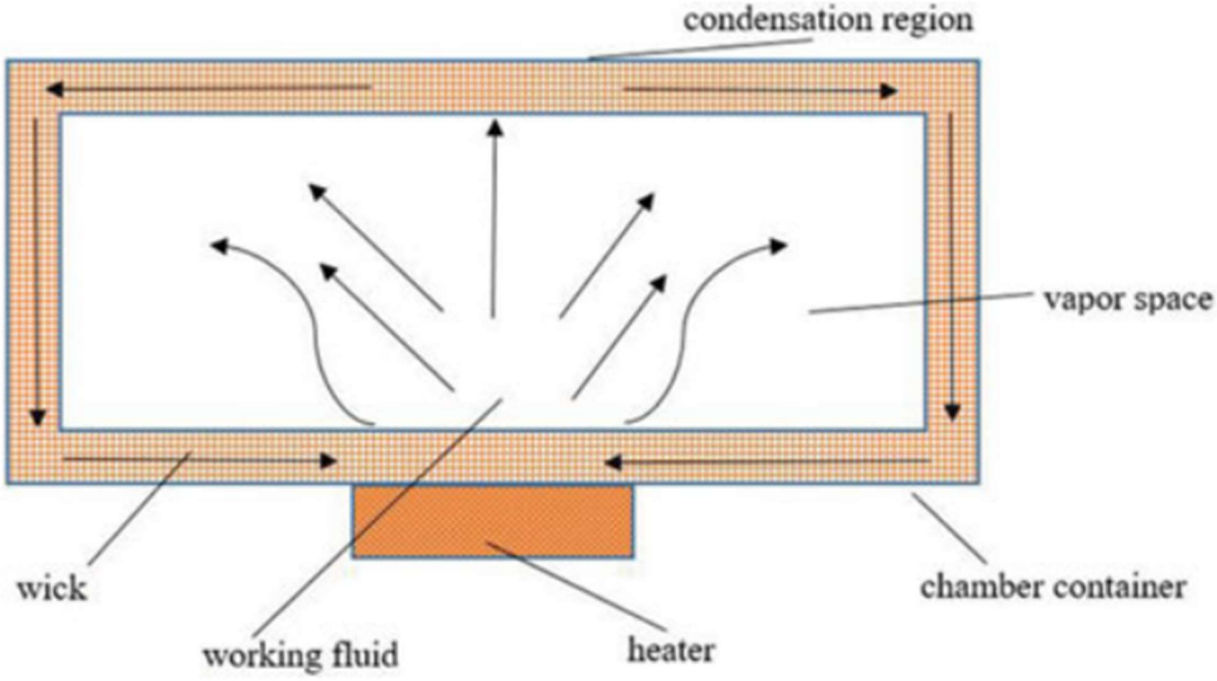


Figure 62 Schematic of Vapor Chamber

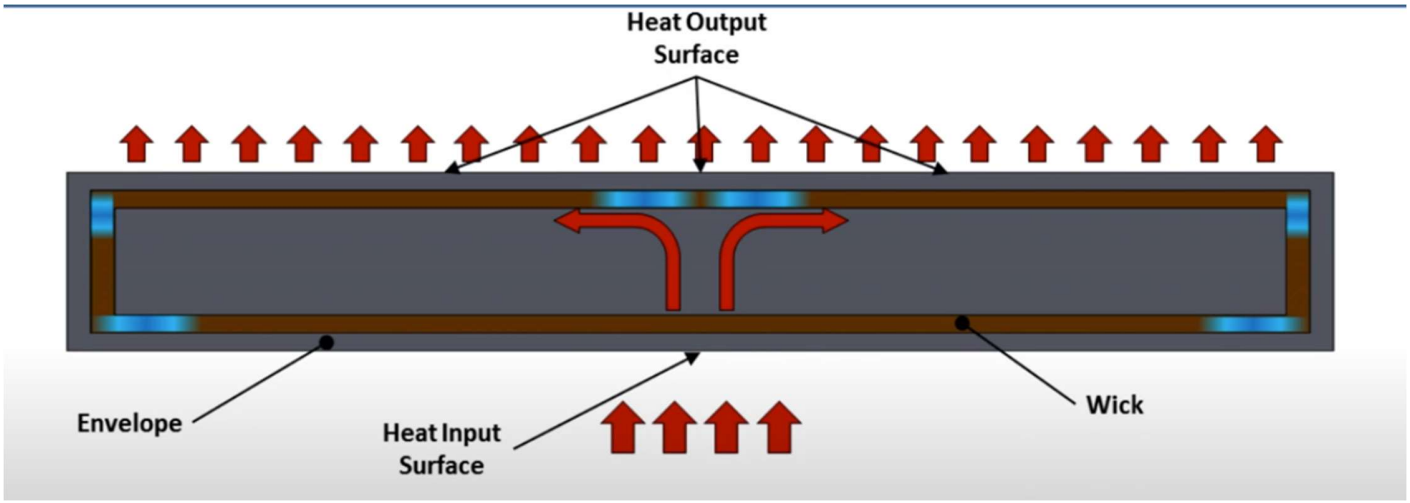


Figure 63 Working of Vapour Chamber

Vapour chamber is used to:

1. Conduct heat to outside heat sink base efficiently.
2. Lower overall temperatures, including hotspots.
3. Control temperature distribution and gradient.

	Vapor Chamber	Heat Pipe
Theory	Two-phase heat transfer	Two-phase heat transfer
Application	Two-dimensional heat distribution. Spreading heat by a single vapor chamber. Suitable for large heat flux and high power	One-dimensional heat distribution. Using one or more heat pipes to spread heat. Suitable for long distance between heat source and heat exchanger.
Shape	Complex shape in x and y direction with pedestal, flattened rectangle, surface embossing and z-direction bendable	Round, flattened or bent in any direction
Sealing method	Swaged and welded on each end	Pinched and welded on each end, four sided diffusion bonding or welding
Fixtures	Mounted with through-holes in vapor chamber	Additional fixture plates needed to mount heat pipes
Mounting to heat source contact	Direct contact	Indirect contact through base plate unless flattened and machined

Figure 64 Comparison between Vapor Chamber and Heat Pipe

Material	Density	Spreading	Thermal Conductivity W/m K	Max. Heat Flux, W/cm^2	Minimum Thickness	Max. Height, cm	Direct Die Attach
Aluminum	1	2-D	200	Depends on Geometry	Structural Considerations	N.A.	N
Spot Heat Pipe	~1.3	1-D	10,000 to 100,000	75 (500)	3 mm (< 1.8 mm flattened)	~ 25 cm (10 in.)	N
HiK™ Plate	0.98-1.2	Hybrid 1-2D	600-1,200	75	1.83 mm (0.072 in.)	~ 50 cm (20 in.)	Y
Vapor Chamber	~2.8	2-D	5,000 to 100,000	1000 (1 cm^2)	3.0 mm (0.120 in.)	15 cm (6 in.)	Y

Figure 65 Comparison of various materials with respect to multiple parameters

- Hence we can conclude that all the above cooling techniques can be applied for thermal management of chiplets in automobiles because they are well tried and tested to work in environments similar to those observed in automobiles.
- Though we would suggest Vapour Chamber considering their better efficiency than heat pipes and better packaging in 2.5D systems considering their structure.

THERMAL MANAGEMENT SIMULATION

Advanced thermal management strategies for high-performance chiplets

The ever-increasing demand for high-performance chiplets in modern electronic systems necessitates a thorough understanding of thermal management strategies to ensure optimal functionality and reliability. In this report, our focus is on exploring advanced thermal management techniques for chiplets, with a particular emphasis on simulation using ANSYS Icepak. ANSYS Icepak is a state-of-the-art simulation tool renowned for its capabilities in predicting and analyzing the thermal behavior of electronic systems. By leveraging finite element analysis (FEA) and computational fluid dynamics (CFD), Icepak enables detailed simulations of heat generation, transfer, and dissipation within chiplet architectures. The utilization of ANSYS Icepak provides us with a robust platform to investigate the intricate thermal challenges associated with densely packed chiplets, offering insights into the effectiveness of various cooling strategies. This report aims to delve into the complexities of chiplet thermal management, supported by simulations conducted with the powerful and versatile ANSYS Icepak software.

HEAT GENERATION IN CHIPLETS

2.5D Packaging:

Thermal Dissipation Simulation in 2.5D Packaging with 4-HBM2 and GPU Integration

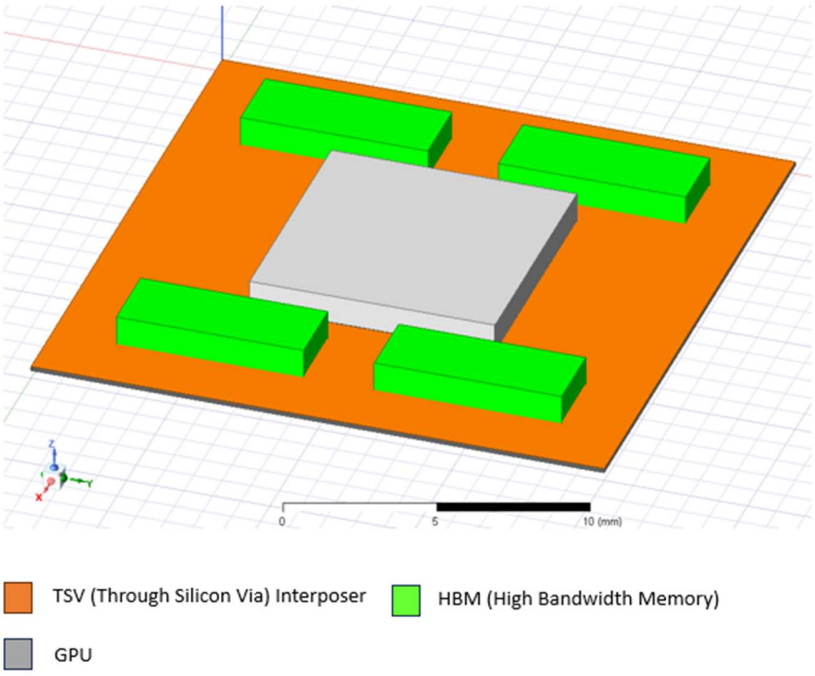


Figure 66 Thermal Dissipation with 4 HBM and GPU

Core Power consumption specifications

HBM2 (maximum memory bandwidth (256GB/s), Bit Rate(2.1Gbps))- 30W
GPU (NVIDIA A100 PCIe)- 250W, ~100% utilization, maximum memory bandwidth (2TB/s)

2.5D packaging using FCBGA (Flip chip Ball Grid Array)

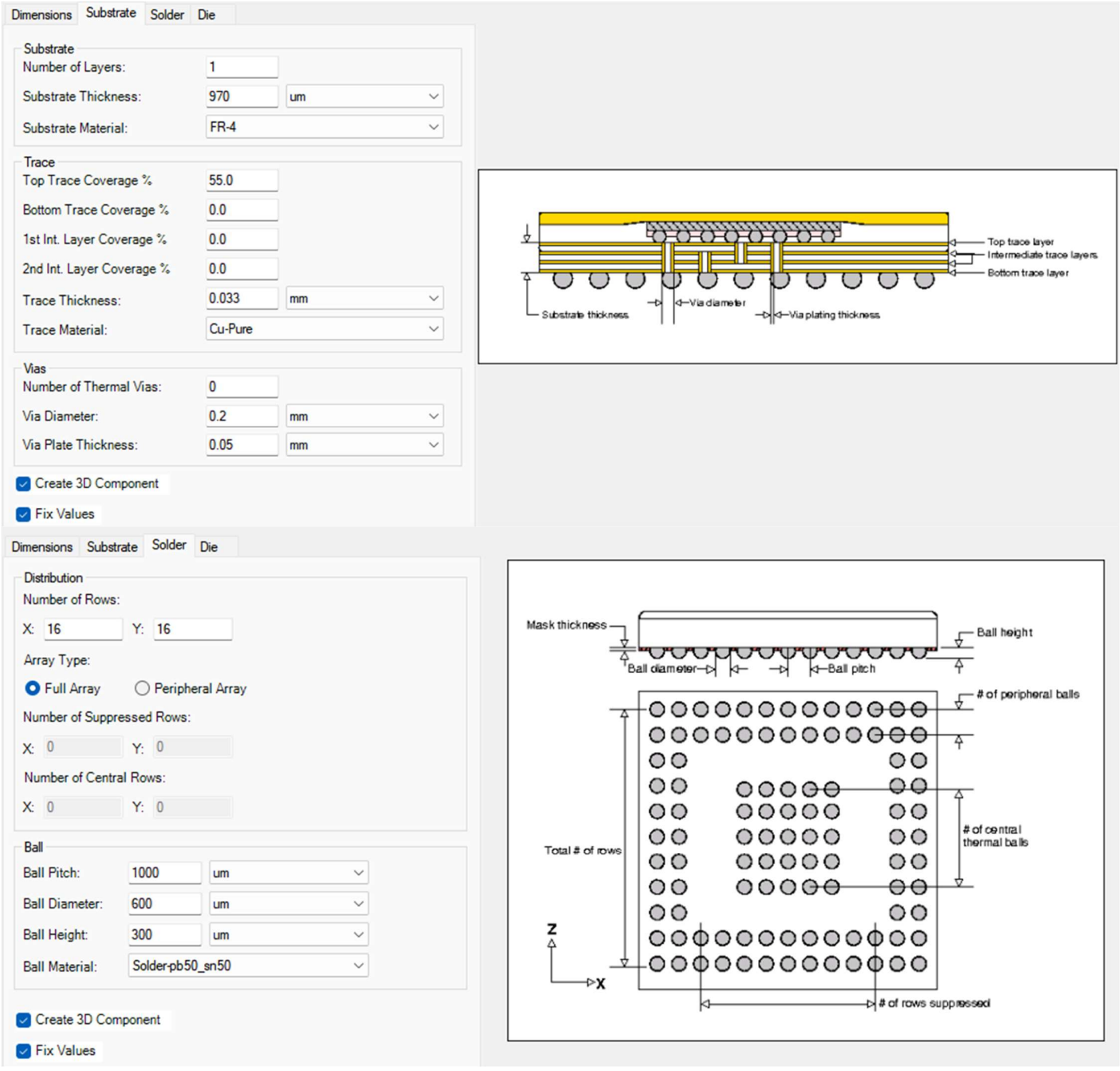


Figure 67 2.5D using FCBGA- Simulation Dashboard

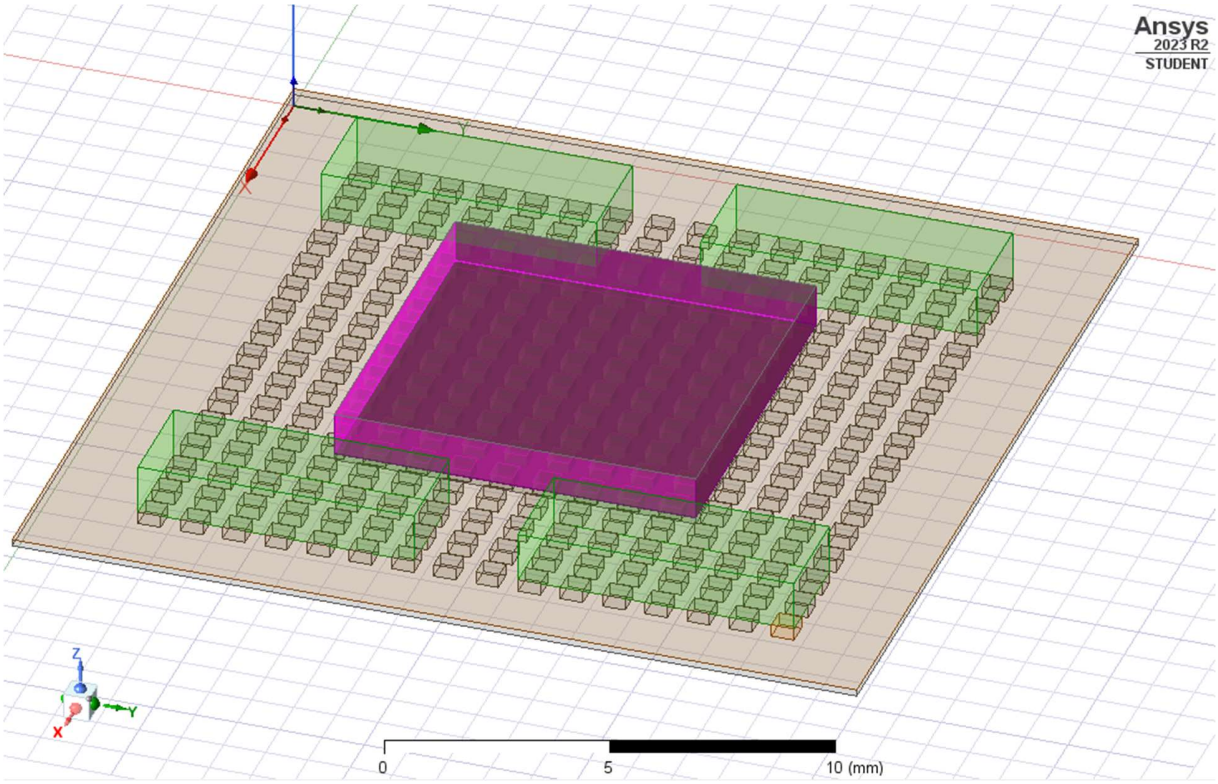


Figure 68 Simulation Screen

Thermal Dissipation (without cooling)

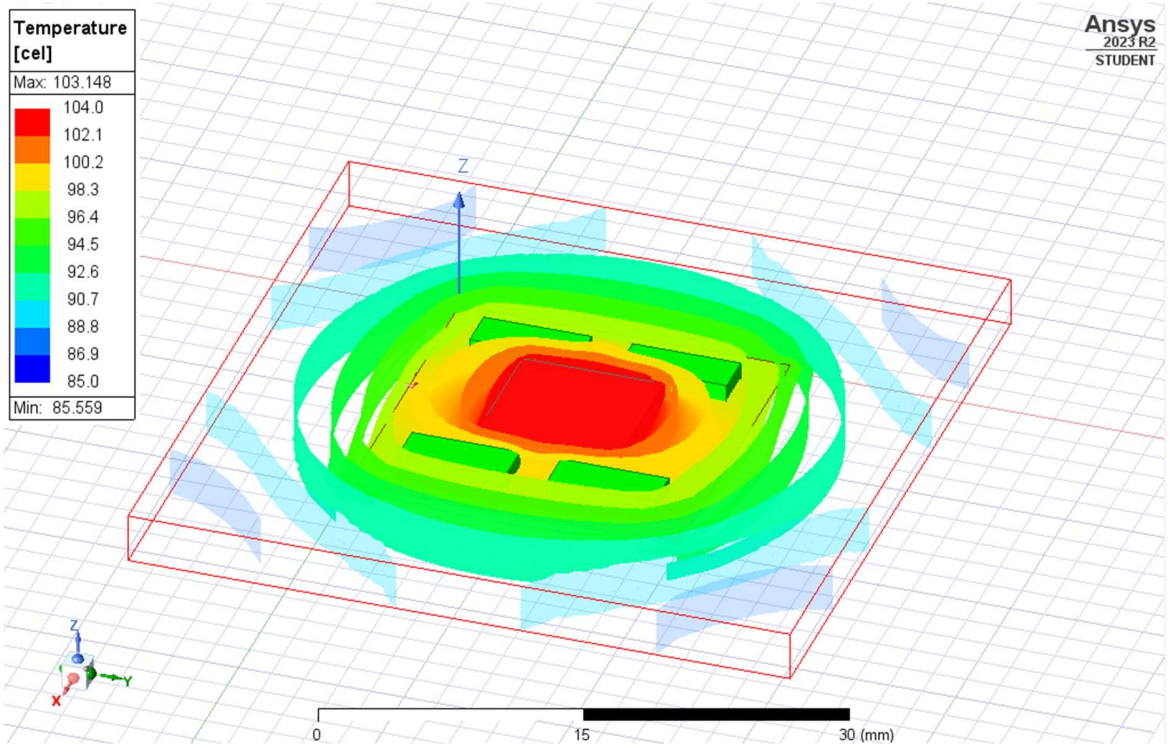


Figure 69 Thermal Dissipation without cooling

Thermal Dissipation (with conventional cooling)

Fan speed 3000 RPM and Aluminum heatsinks

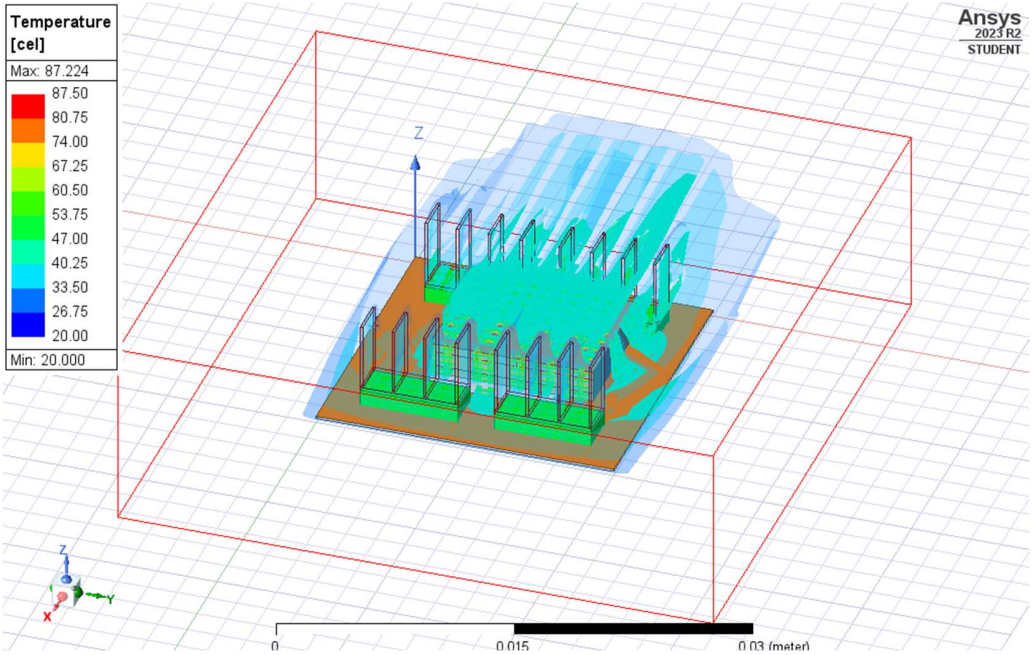


Figure 70 Thermal Dissipation with Thermal Cooling

Thermal Dissipation (Undervolting within Guardband region Task 2.2C)

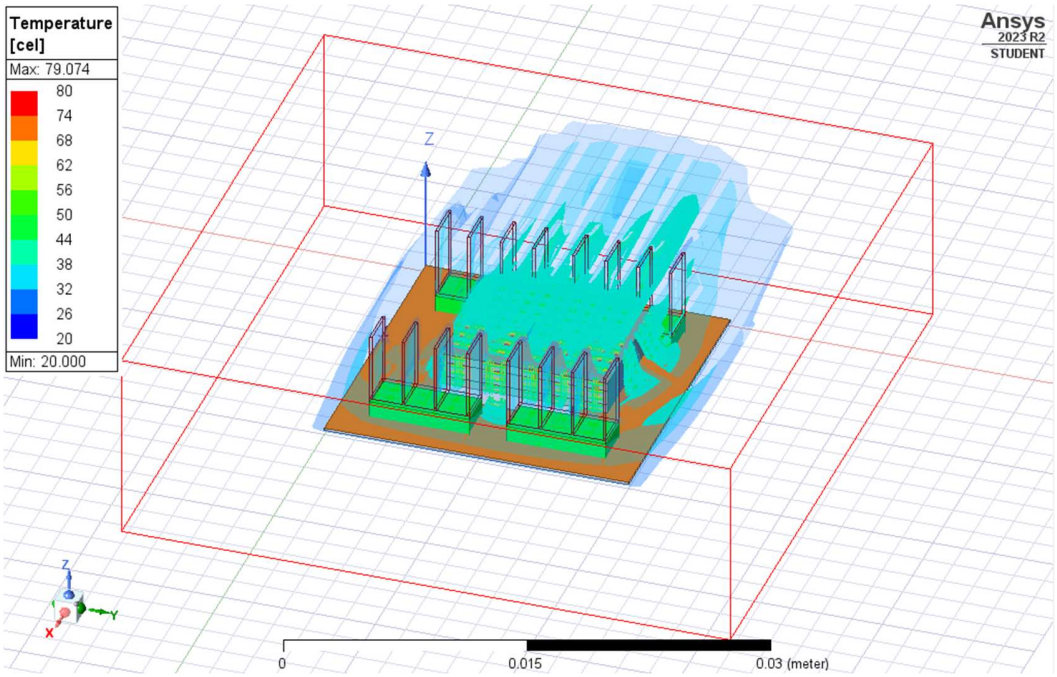


Figure 71 Undervolting within Guardband region

Temperature range - 20 ° - 79.074 ° Celsius

$$Power = Vdd2 * C * f * \alpha$$

The active power usage of a DRAM chip correlates directly with the square of the supply voltage (Vdd). C represents the active load capacitance, f stands for the operating frequency, and α denotes the activity factor, determining the average charge/discharge rate of the capacitor. Consequently, undervolting is anticipated to result in a quadratic decrease in active power consumption.

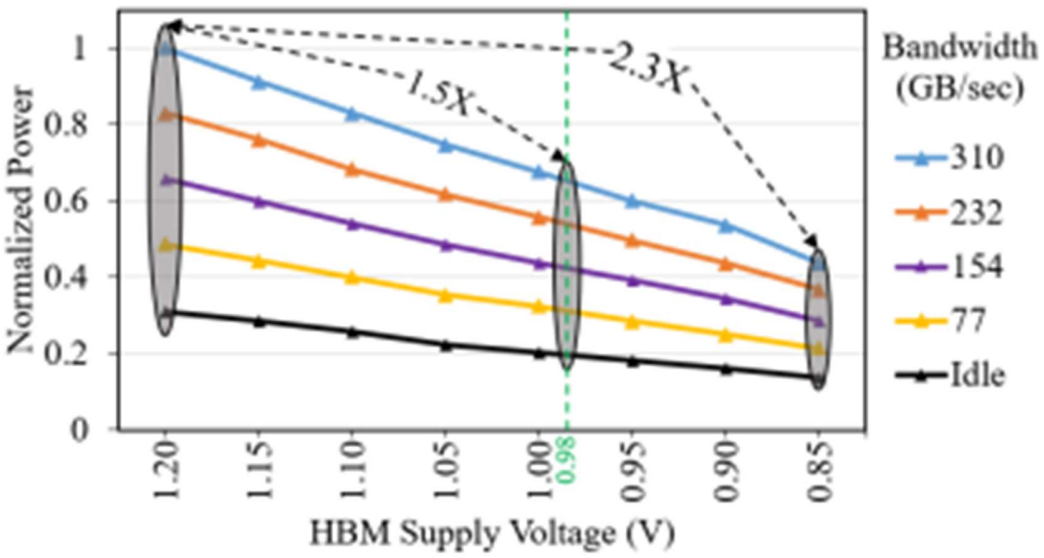
This passage describes two distinct regions in the context of supply voltage for HBM (High Bandwidth Memory):

GUARDBAND REGION:

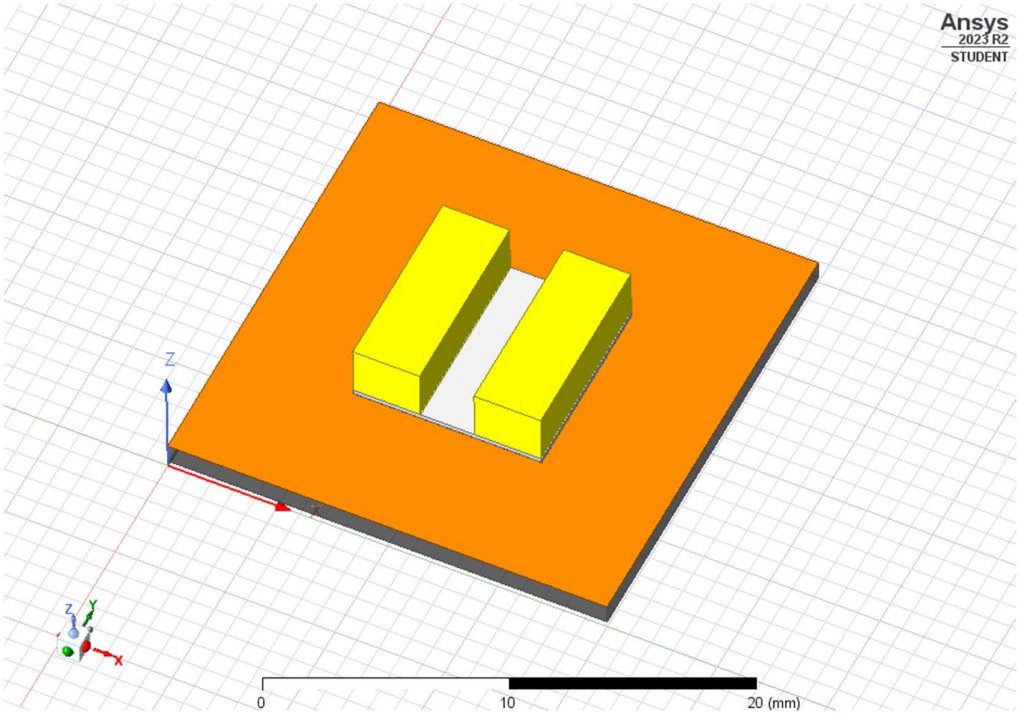
The guardband region extends from the nominal voltage (Vnom = 1.2V) to the minimum safe voltage (Vmin = 0.98V). Within this voltage range, no memory faults are detected. It is considered a secure zone for all operations and workloads, making it crucial for applications intolerant to memory faults to function within this designated voltage range.

UNSAFE REGION:

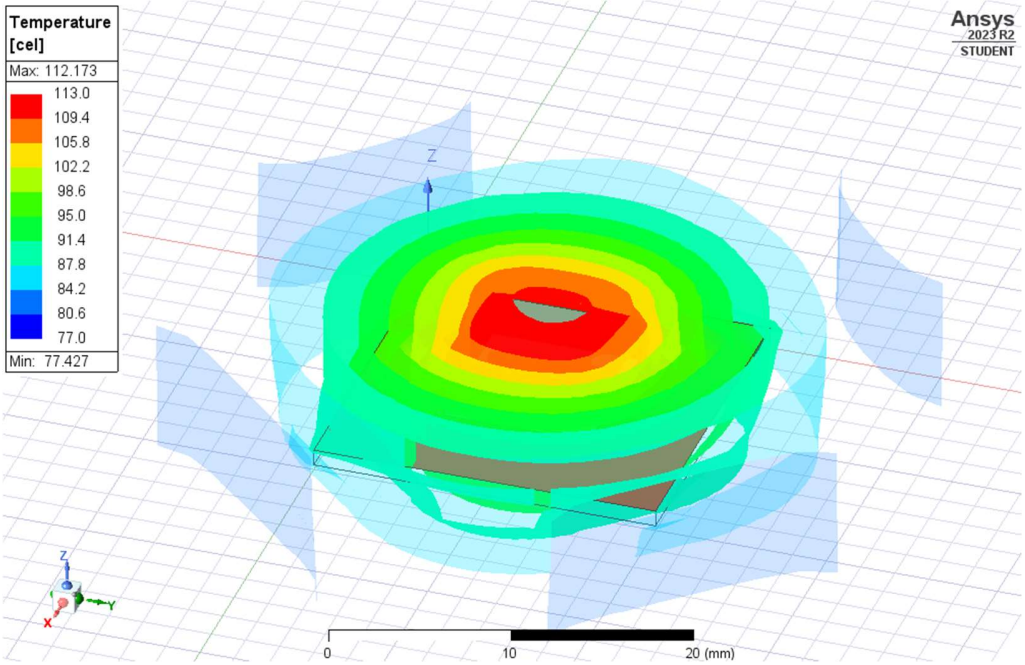
The unsafe region encompasses voltages lower than Vmin. Within this range, faults manifest, and any application employing HBM with supply voltages in this unsafe region should assess the consequences of these faults to guarantee proper functionality. Decreasing the voltage introduces additional faults, displaying an exponential increase in occurrence until approximately 0.84V, where all memory bits undergo 0-to-1 or 1-to-0 bit flips. Further reduction below 0.84V, down to the minimum working voltage Vcritical = 0.81V, renders the entire HBM components faulty.



3D packaging of GPU and 2- HBM2 using FCBGA

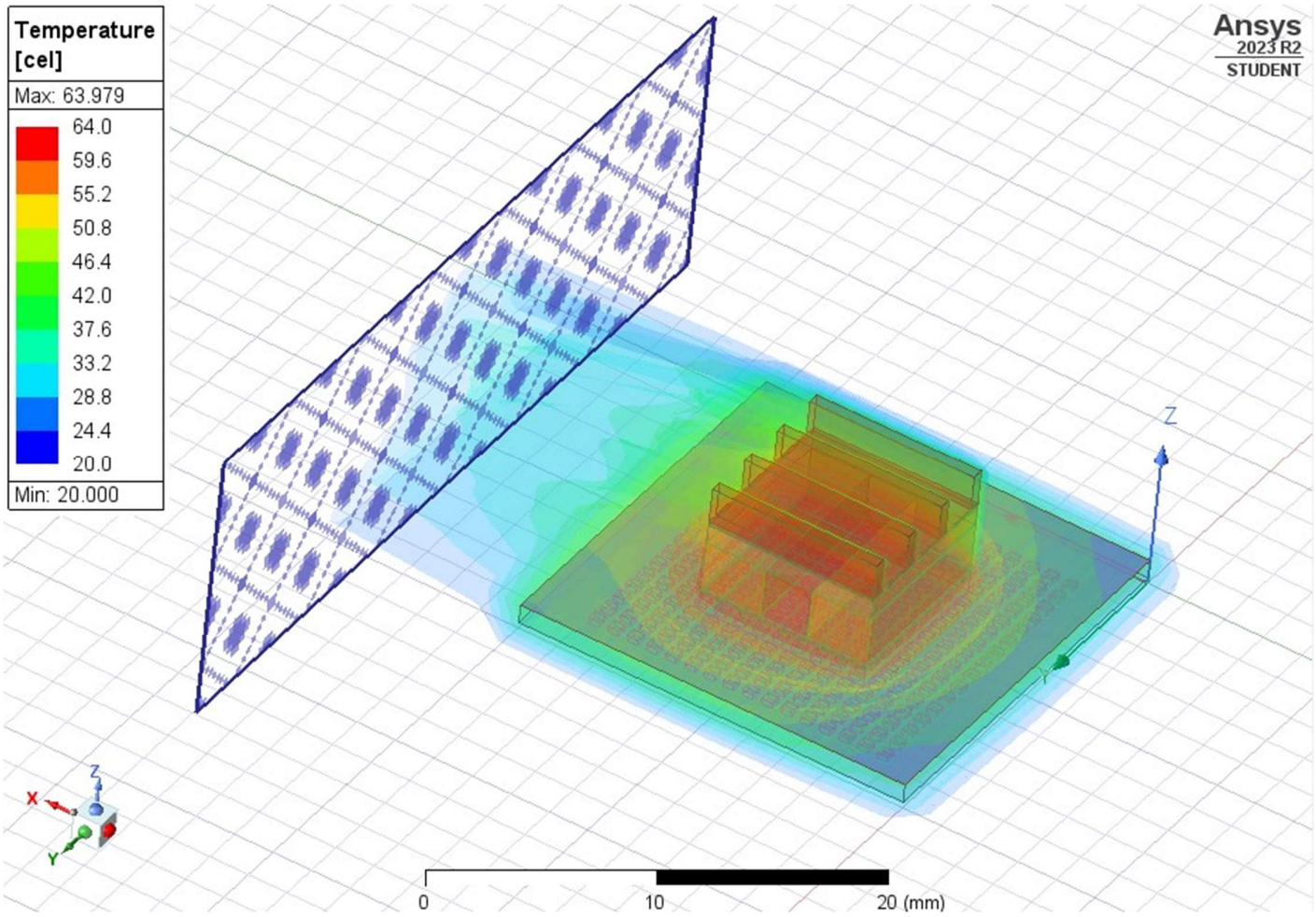


Thermal Dissipation (without cooling)



Thermal Dissipation

Fan speed 3000 RPM, Aluminium heatsink and undervolting



Temperature range – 20 ° - 64 ° deg Celsius

Thermal issues in 3D packaging of chips are significant considerations due to the increased power density and reduced spatial distribution in vertically stacked layers. Here are some key thermal challenges associated with 3D chip packaging:

1. Heat Dissipation:
    a. Limited Surface Area: In 3D packaging, the available surface area for heat dissipation may be constrained, making it challenging to efficiently transfer heat away from the stacked chips.
    b. Vertical Heat Flow: Heat generated in one layer may affect the performance of the layers above or below, leading to potential thermal bottlenecks.
2. Interlayer Thermal Coupling:

c. Thermal Resistance Between Layers: The presence of thermal interfaces between stacked layers introduces additional thermal resistance, impacting the overall heat dissipation efficiency.
d. Thermal Crosstalk:Thermal interactions between adjacent layers can result in uneven temperature distribution, potentially affecting the reliability and performance of the entire package.

Thus use of Thermal fillers inbetween chiplets can result into better cooling.

Thermal Dissipation

Fan speed 3000 RPM, Aluminium heatsink, undervolting and Aluminium Nitride filler in between HBM2 blocks.

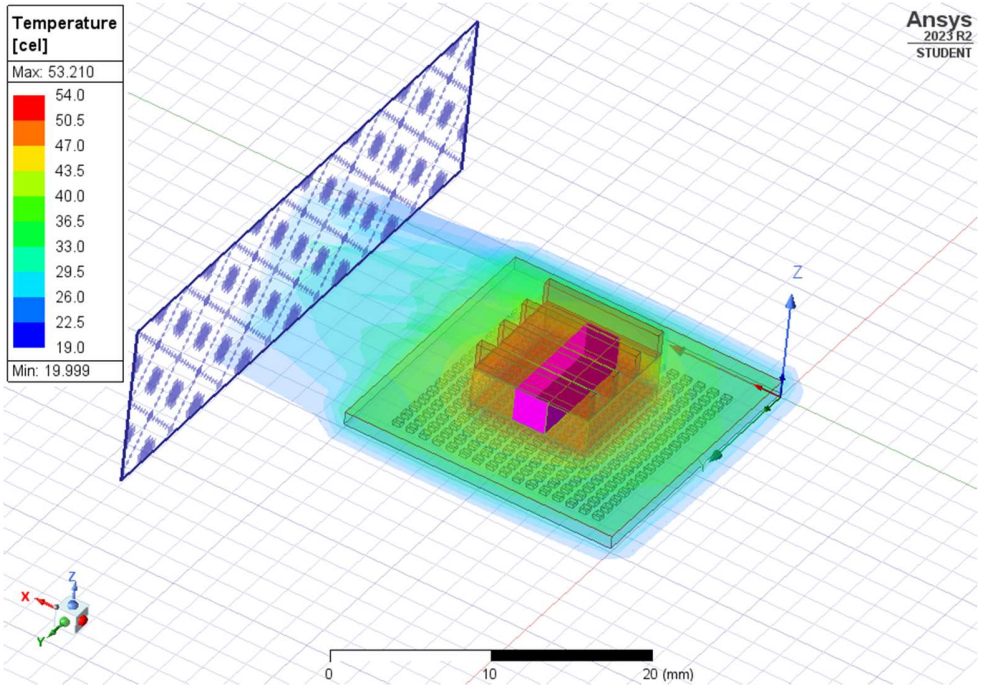


Figure 72 Fan speed 3000 RPM, Aluminium heatsink, undervolting and Aluminium Nitride filler in between HBM2 blocks.

Temperature range:

With cooling technology - 19 ° - 54 ° Celsius
Without cooling technology – 77 ° – 112 ° Celsius

2.5D Design for Chiplet on Interposer in CADENCE SIP

Below is described the placement and routing of Chiplets on Interposer designed in the Cadence SiP. Command line scripting has been utilized for the chiplet design and setting up of the design parameters.

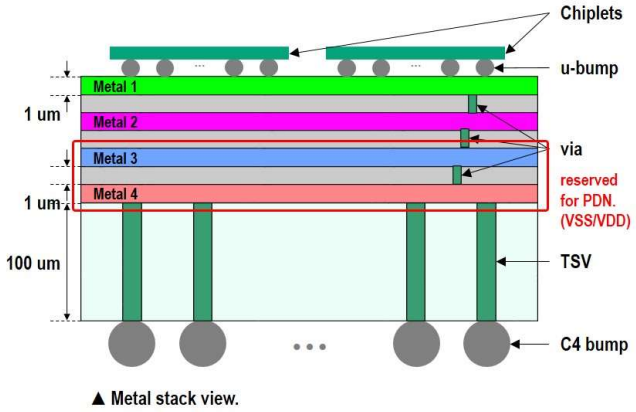


Figure 73 Schematic of the Design

The detail of the interposer technology used is:

- 65nm silicon interposer
- metal layer #: 4
- metal w/s: 0.4um/0.4um
- u-bump diameter/pitch: 20um/40um
- C4 bump diameter/pitch: 90um/180um

The model dependencies are Cadence SiP Layout XL 17.2 and Cadence PCB Automatic Router. A techfile has been generated for the design that contains all the necessary information, such as process technology, design specifications, material properties and metal counts.

```
metal_count=4           # the number of metal layers.
metal_thickness=1       # unit: um
dielectric_thickness=1      # unit: um
substrate_thickness=100     # unit: um
wire_width=0.40         # unit: um.
wire_spacing=0.40
```

SiP script has been generated from the script make_sip_script for importing the techfile, chiplet placement and grid setting.

The other script files are for PDN generation, VSS and VDD. A good power delivery network(PDN) is responsible for efficiently distributing power across the design, ensuring stable voltage levels, and

minimizing voltage drops. The VSS represents ground connection, and the VDD represents positive power supply. The VSS and VDD routing is for pathways for ground and power connections. For this design, automatic routing and meshing have been done.

The Fanout distributes signals from a single source from 'via'. This fanout helps optimize routing efficiency by reducing congestion around a specific point in the layout. It has been created on the interposer with the following features:

- Start layer: M4
- End Layer: UBM
- Via has been used as TSV

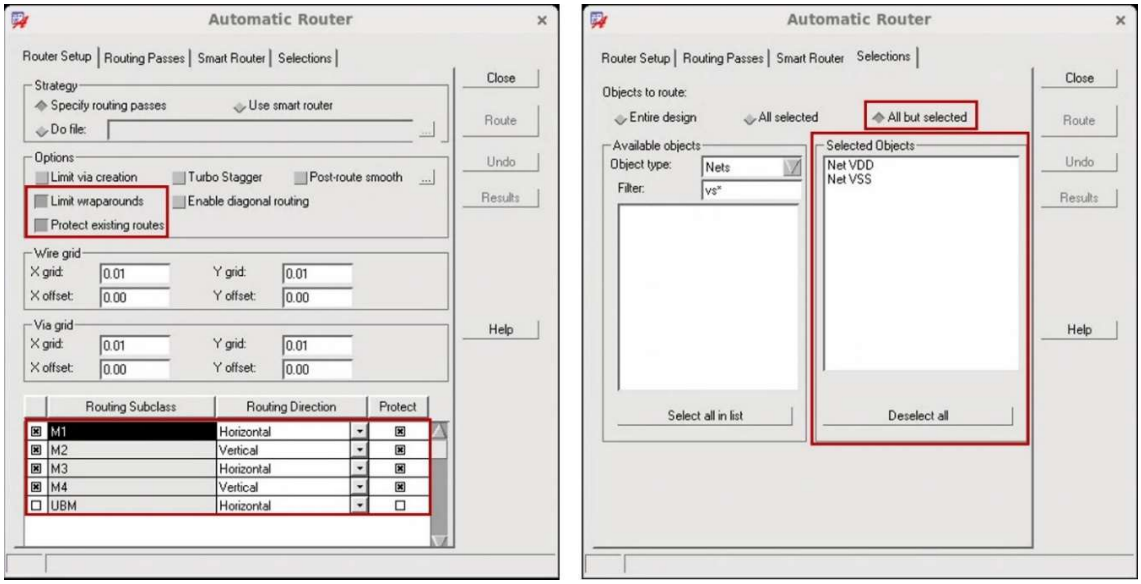


Figure 74 Automatic router setting

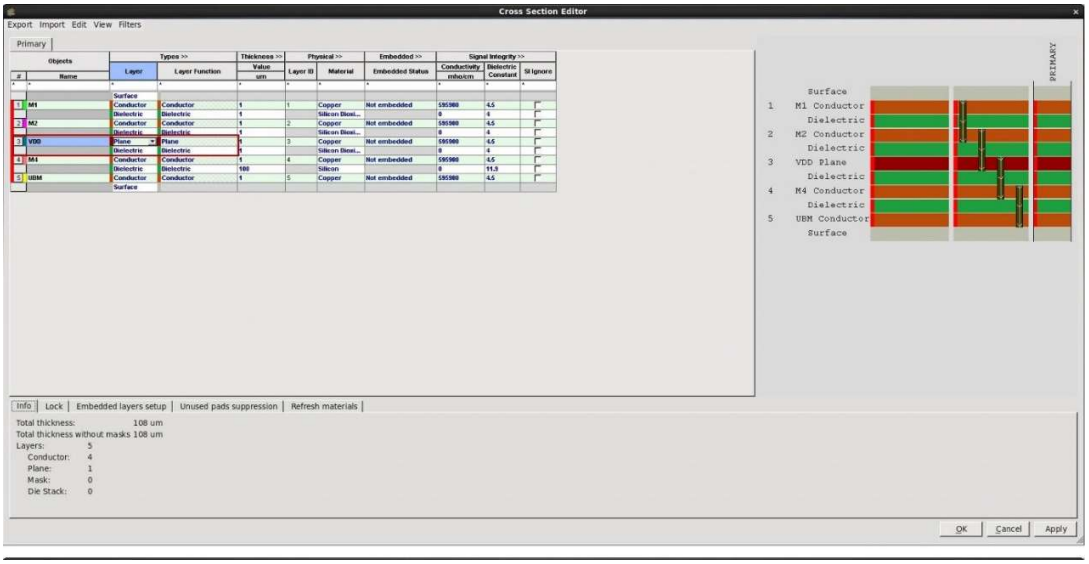


Figure 75 VDD Routing Setup

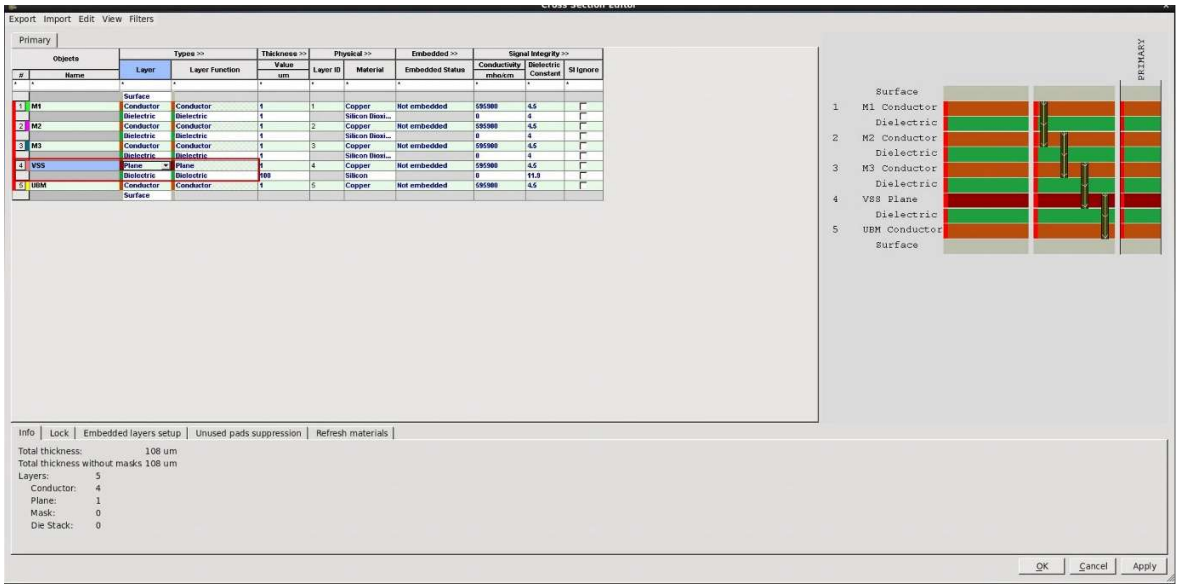


Figure 76 VSS Routing Setup

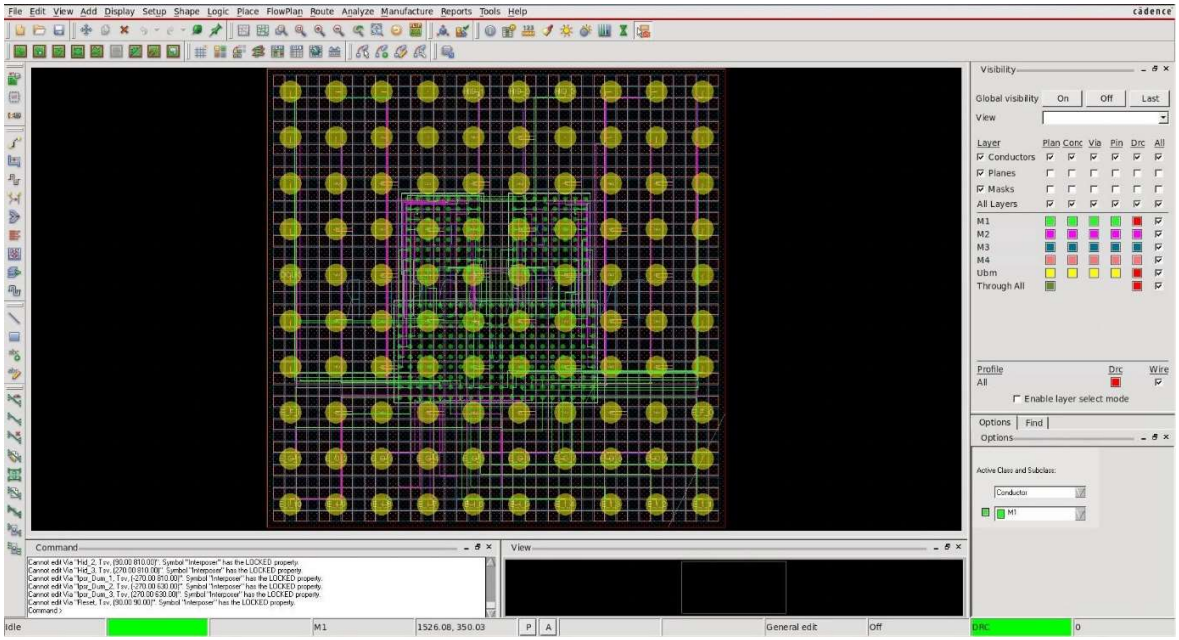


Figure 77 Signal Routing Result

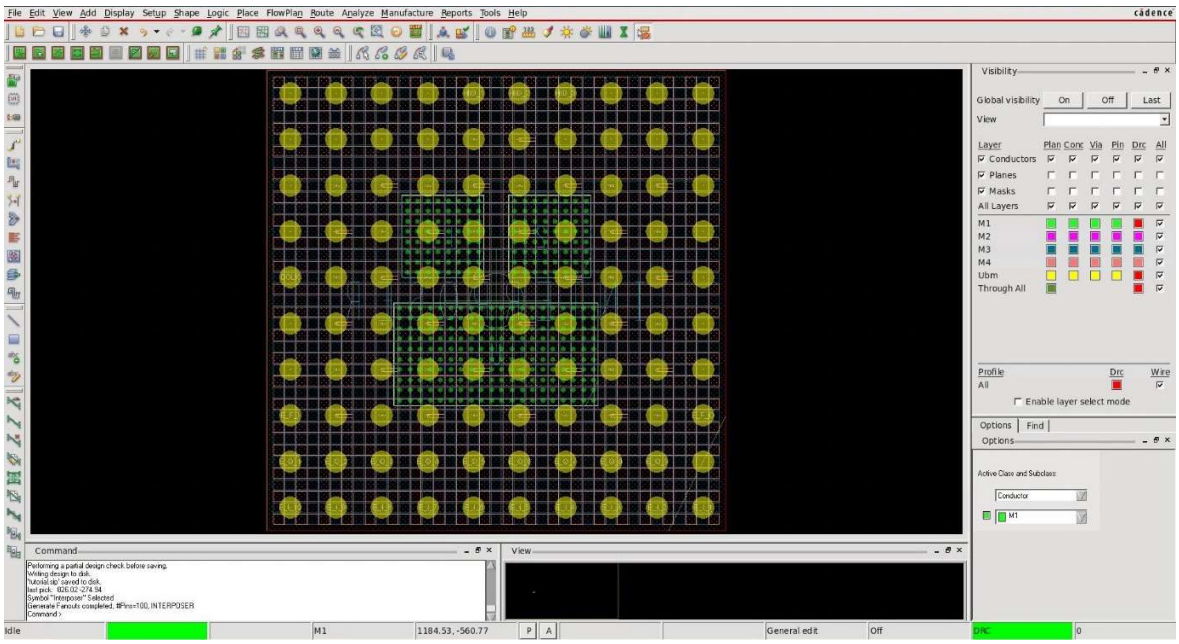


Figure 78 Fan-out and PDN generation result

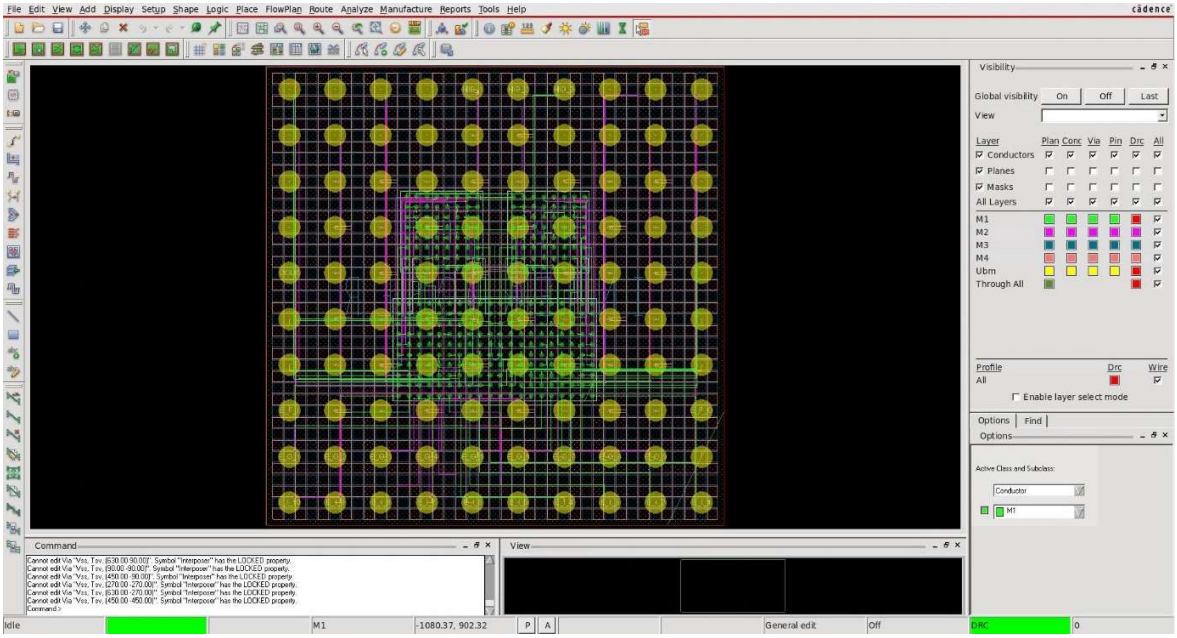


Figure 79 Final Interposer Placing and Routing

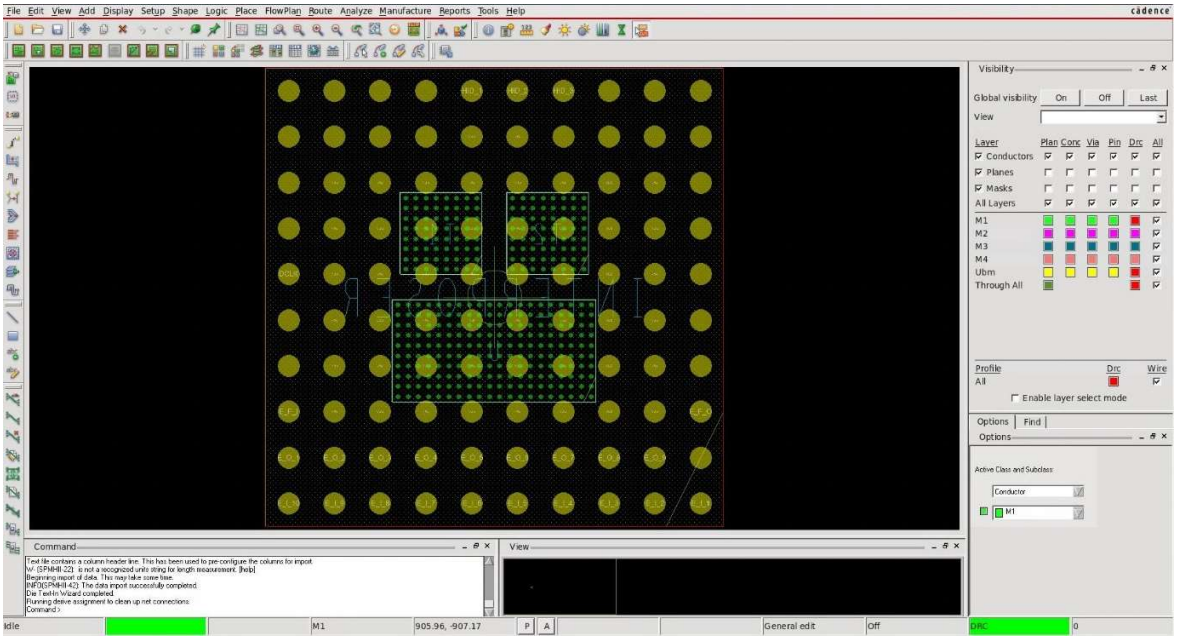


Figure 80 Chiplet Placement Result

We have demonstrated a basic 2.5D packaging with two chiplets interconnected on an interposer with four metal plates. This design considers U bump and C4 bumps, TSVs and 'via' reserved for power delivery networks.

INNOVATIVE TECHNOLOGY

NEURAL PROCESSING UNIT

Neural Processing Units (NPUs) are dedicated hardware accelerators designed to perform the intensive computations required by neural networks efficiently. Neural networks, especially deep neural networks used in machine learning and AI, involve complex mathematical operations and large-scale matrix multiplications. General-purpose processors, such as CPUs and GPUs, may struggle to handle these computations efficiently due to their inherent design, leading to high power consumption and slower processing speeds.

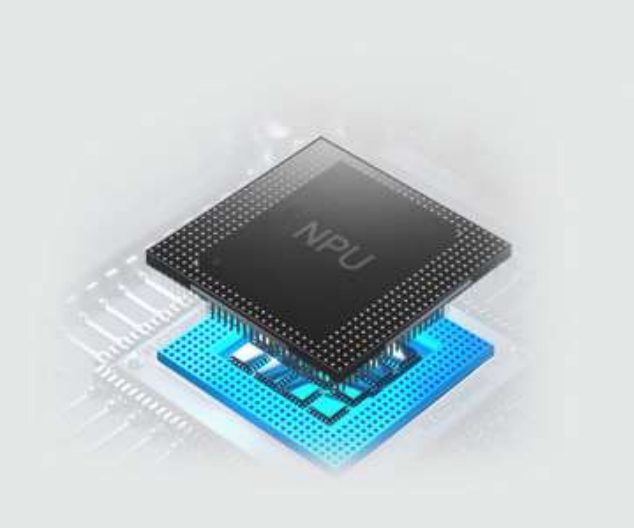


Figure 81 Typical NPU

NPUs are purpose-built to excel in the specific types of calculations involved in neural network processing. These units often feature highly parallel architectures, optimized memory hierarchies, and specialized instruction sets tailored for neural network workloads. The goal is to provide faster and more power-efficient execution of AI models.

WHY NPU IS CRUCIAL FOR AUTONOMOUS CARS:

Real-time Processing:

- Autonomous cars require real-time decision-making capabilities to interpret data from various sensors, such as cameras, lidar, radar, and ultrasonic sensors.
- NPUs excel at parallel processing and are capable of executing neural network inferences quickly, allowing the car to make split-second decisions in dynamic environments.

Complex AI Models:

- The AI models used in autonomous driving are often complex, involving convolutional neural networks (CNNs) for image recognition, recurrent neural networks (RNNs) for sequential data, and other sophisticated architectures.
- NPUs are optimized for handling the intricate computations involved in these models, ensuring efficient execution and high throughput.

Power Efficiency:

- Automotive systems, including those in autonomous cars, have stringent power constraints. NPUs are designed to provide high performance with low power consumption, making them suitable for embedded systems in vehicles.

Edge Computing:

- Autonomous cars often rely on edge computing, where computations are performed locally on the device rather than relying on distant servers. NPUs are well-suited for edge processing, reducing latency and enhancing the system's responsiveness.

Sensor Fusion:

- Autonomous vehicles gather data from multiple sensors simultaneously. NPUs support sensor fusion by efficiently processing and integrating data from various sources, enabling a comprehensive understanding of the vehicle's surroundings.

Adaptability:

- NPUs can be programmed or configured to adapt to evolving AI models and changing environmental conditions. This adaptability is crucial in the dynamic scenarios encountered in autonomous driving.

Safety and Redundancy:

- Redundancy and safety are paramount in autonomous systems. NPUs contribute to redundancy by providing dedicated processing for critical AI tasks, enhancing the overall reliability of autonomous driving systems.

AI CHIP TECHNICAL DETAILS

InferX AI

Category	Details
Product Name	InferX™ Vision AI IP+SW
Target SoC	Orin AGX or better in your SoC
Inference Performance	Fast vision AI inference at low cost, low power
Hardware Configuration	- 80% hardwired, 100% reconfigurable (GPU-like performance)
	- TSMC Nodes: 16, 7, 5, 4, and 3nm
	- 10's of square millimeters
	- 10x cheaper, 10x lower power, and smaller than Orin AGX with less DRAM BW
Model Compatibility	- Supports super-resolution models like Yolov5L6 (1280x1280 pixels at 30 frames/second)
	- Optimized for batch=1 and very high accuracy
	- Runs Transformers efficiently and accurately
InferX Compiler	- Converts high-level neural network models to InferX code
	- Supports multiple models simultaneously
	- Beta version available for evaluation under Software License
InferX Inference Compiler	- Takes in neural network models in high-level formats
	- Converts to INT8 for high performance with accuracy within <1% of FP16's mAP
	- Automatic quantization; no shortcuts on precision
InferX Hardware	- Tensor Processing Tile architecture
	- Scalable and reconfigurable for optimization
	- eFPGA core for adaptation to new operators & activations
	- Runs super-resolution models on megapixel images (N7, batch=1)
Tensor Processing Tile (TPU)	- 16 hard-wired 1-dimensional tensor processors
	- Capable of processing in INT8, INT16, and BF16 modes
	- Programmable interconnect for configuration
	- Reconfiguration takes a few microseconds
	- 16 Dense TOPs (N5) for efficient utilization and minimal DRAM bandwidth

InferX Arrays	- Delivered as an array of tiles from 1x1 to any desired size
	- AXI bus interface for SoC connection
	- Linear performance scaling: N times faster for an array with N tiles
Temperature and Design Features	- AXI interface, -40C to +125C design
	- Very high DFT coverage, full test vectors
InferX Compiler Features	- High precision with no pruning or model modifications
	- Beta version available for evaluation under Software License
TSMC IP Alliance Membership	- TSMC IP Alliance Member
	- Compliance for TSMC9000 design methodology, validation, and documentation
	- Implemented IP in various TSMC nodes: 40, 28, 16, 12, 7, and started on 5nm
About Flex Logix	- eFPGA licensed for 40 chips with >20 working in Silicon

Conclusion:

High Performance and Efficiency: The InferX™ Vision AI boasts GPU-like performance with 80% hardwired capabilities and is optimized for low cost, low power, and small form factor.
Adaptability: The inclusion of an eFPGA core allows for adaptability to new operators and activations, ensuring flexibility for evolving AI models.

HAILO AI

Category	Details
Processor Model	Hailo-15™ AI Vision Processor
AI Performance	Up to 20 TOPS (Tera Operations Per Second)
	Powerful Neural Network (NN) Core for processing multiple advanced DL models in parallel
Deep Learning Model Processing	High FPS deep learning model processing for faster and highly accurate detection of more objects per frame
Power Consumption	Best AI performance at a standard camera power consumption & cost envelope
Software and Frameworks	Industry-standard frameworks with a complete Yocto-based Linux distribution
SoC Interfaces	Variety of interfaces for image sensors, data, and memory
Security Features	Secure boot and secure debug with hardware-accelerated crypto library, TrustZone, TRNG, and Firewall
Image Quality Enhancement	AI-powered vision processing for video image enhancement: noise reduction, digital zoom, image stabilization, dynamic range, and distortion correction

ISP and Vision Sub-system	Premium 4k60 image quality with an advanced ISP pipeline and Vision sub-system
	DSP vision processing for high-quality video encoding
Sensor Interfaces	Video In: dual MIPI CSI, DVP 24 bit
	Video Out: MIPI DSI
Peripheral Interfaces	PCIe Gen 3.0 x 4 lanes (Endpoint or RC)
	10/100/1000 Ethernet with RMII/RGMII
	USB 3.1 Gen2 Host/Device, 2.0 Host
Application Processor Sub-system	Quad-core ARM™ A53 up to 1.3 GHz, 12 kDMIPs
DSP Vision Processing Sub-system	Vector DSP, 256 MACs @ 700 MHz, supports up to 350 GOPs
Memory Interfaces	LPDDR4/4X 32bit @4266 MT/s
	QSPI
	SDIO 3.0/eMMc 5.1 (up to HS200)
Security	Secure boot, secure debug, hardware-accelerated crypto library, TrustZone, TRNG, Firewall
Physical Specifications	Packaging: FCCSP 15x15 mm
	Operating temperature: -40°C to 85°C
Vision Sub-system	ISP: up to 12MP resolution, 600 Mpixel/s pixel rate
	RGGB & RCCB CFAs support
	Up to 3 exposures merging
	WDR for low light image processing
	Advanced noise reduction features: 2DNR, 3DNR, Chroma NR
Video Encoding	HEVC & AVC (H.265/H.264), multiple stream
	Image Stabilization, Lens Shading & Distortion Correction, Digital Zoom, Flip & Rotate
Comprehensive Software Package	Drivers, libraries, and tools for smart cameras development
Camera Development Kit	Hardware evaluation platform with integrated Hailo-15™, software, documentation, and support
Performance Levels	- Hailo-15H: 20 TOPS
	- Hailo-15M: 11 TOPS
	- Hailo-15L: 7 TOPS

Conclusion:

- **Efficiency at Standard Power: Hailo-15™** provides high AI performance within a standard camera power consumption envelope, making it suitable for various applications.
- **Versatility:** The processor supports diverse use cases such as voice interface, speaker identification, and multi-sensor fusion, showcasing its adaptability.

MLSoC AI

Category	Details
Product Name	SiMa.ai MLSoC
MLSoC Architecture	Software-centric approach for effortless machine learning deployment at the embedded edge ML market
Application Development	- Supports legacy applications and future ML use cases
	- High performance, low power, safe, and secure ML inferencing
	- Optimization, visualization, and debug tools
	- Model Zoo for ready-to-use DNN models
	- Ability to run up to four distinct DNN models concurrently
Framework Compatibility	- Optimizes and supports DNN models from TensorFlow, PyTorch, ONNX, MXNet, etc.
MLSoC Performance	- Up to 50 TOPS (Tera Operations Per Second) for neural network computation
Form Factor	- FCBGA 1369 balls; 31mmx31mm, 0.8mm pitch
Target Applications	- Smart Vision
	- Robotics and Industry 4.0
	- Drones
	- Automotive
	- Healthcare
	- Government Sector
Processing System Components	- Machine Learning Accelerator (MLA) providing 50 TOPS at 10 TOPS/W
	- Application Processing Unit (APU) - Cluster of four Arm Cortex-A65 processors
	- Video Encoder/Decoder supporting H.264 and HEVC standards
	- Computer Vision Unit (CVU) with a four-core Synopsys ARC EV74 video processor
	- High-speed I/O Subsystem with 4 Gigabit Ethernet ports and PCIe Gen4 8-lane interface
	- Low-speed I/O Subsystem with interfaces like SPI, I2C, GPIO, and SDIO/eMMC
	- DRAM Interface System (DIS) supporting four 32-bit LPDDR4 memory controllers
	- Boot and Security Unit (BSU) for secure key storage, decryption, authentication, and security API
Software-First Development	- Co-designed software toolchain and hardware for optimized performance
	- Intermediate representations, including TVM Relay IR
	- Compiler optimization techniques for supporting a wide range of frameworks

Conclusion:

- **Versatile ML Deployment:** MLSoc is a purpose-built MLSoC platform supporting legacy and future ML use cases, optimizing DNN models from various frameworks with high performance and low power.
- **Application Flexibility:** It offers the ability to run up to four distinct DNN models concurrently, providing flexibility for complex AI applications.

NDP120 AI

Category	Details
Product Name	Syntiant® NDP120 Neural Decision Processor™
Purpose	Special purpose chip for audio and sensor processing for always-on applications
Architecture	Syntiant Core 2™ programmable deep learning architecture
Deep Learning Support	Natively runs multiple Deep Neural Networks (DNN)
	Supports architectures such as CNN, RNN, and fully connected networks
	Delivers 25x tensor throughput compared to Syntiant Core 1™ in NDP100 and NDP101 devices
Audio Processing	Programmable HiFi 3 DSP available for classical audio processing
	Supports far-field, near-field, and close-talk voice interface
	Supports multiple wake words, local commands, AEC, noise suppression, beamforming
	Speech enhancement, speaker identification and verification, acoustic event and scene classification, multi-sensor fusion
Neural Network Performance	Hardware acceleration support for up to 6.4 GOPS/s
	Supports various neural network layers: fully-connected, 2D convolution, depth-wise convolution, recurrent neural network including LSTM and GRU, average and max pooling
	Concurrent support for multiple neural networks
	Up to 896k neural parameters in 8-bit mode, 1.8M parameters in 4-bit mode, and more than 7M parameters in 1-bit mode
Audio Interfaces	Quad PDM digital microphone interface
	Dual I2S channels or TDM4 streaming interfaces
	Support for up to 7 audio streams including I2S/TDM output audio interface for streaming audio output, including post-processed audio
Peripheral Interfaces	I2C controller and target modes for sensor control and integration
	QSPI target & controller interfaces
	26 GPIO pins
	Input holding-tank with up to 10 seconds of audio recording and faster-than-real-time extraction
	Embedded Arm Cortex-M0 for device management with 48KB SRAM, dual timers, and UART functionality
	Low power PLL for flexible clock input
	Onboard firmware decryption and authentication
Processor Specifications	Up to 100MHz internal operating frequency
Packages	3.1mm x 2.5mm 42-ball WLBGA package (0.4mm pitch)
	5mm x 5mm 40-pin QFN package (0.4mm pitch) - also available as an AEC-Q100 Grade 3 qualified automotive SKU
Power Efficiency	Low power consumption suitable for battery-powered devices and power-constrained systems

SDK and TDK Support	Software Development Kit (SDK) integrates into any software environment
	Training Development Kit (TDK) enables the use of standard frameworks such as TensorFlow for customer-programmed applications
Applications	Enables speech and sensor interfaces in small systems
	Enables always-on detection usage models

Conclusion:

- **Low Power for Always-On Applications:** NDP120 is designed for always-on applications in battery-powered devices, emphasizing low power consumption.
- **Audio and Sensor Processing:** It excels in audio processing for voice interfaces, acoustic echo cancellation, and sensor fusion, making it suitable for a range of applications.

Comparison Table between AI Chip

Feature	InferX™ Vision AI	Hailo-15™ AI Vision Processor	Syntiant NDP120	MLSoC AI
AI Performance (TOPS)	80% hardwired, GPU-like	Up to 20 TOPS	Hardware acceleration support for up to 6.4 GOPS/s	Up to 50 TOPS
Process Technology	TSMC 16, 7, 5, 4, 3nm	Not specified	Not specified	Not specified
Power Efficiency	Not specified	Best AI performance at standard camera power consumption & cost envelope	Low power consumption suitable for battery-powered devices and power-constrained systems	Low power consumption (effortless ML deployment)
Neural Network Support	Multiple models simultaneously, Transformers, Yolov5L6	High FPS deep learning model processing, support for various frameworks	Natively runs multiple DNNs, supports CNN, RNN, fully connected networks	Optimized and supports DNN models from various frameworks
Model Conversion	InferX Compiler	Not specified	SDK integrates into any software environment, TDK for standard frameworks like TensorFlow	Co-designed software toolchain and hardware, supports a wide range of frameworks

Interfaces	AXI, -40C to +125C design	Various interfaces including video and memory	Quad PDM digital microphone interface, I2S, TDM, I2C, QSPI, GPIO, more	PCIe Gen4, 1G Ethernet, USB 3.1, SDIO/eMMC, SPI, I2C, GPIO, etc.
Memory Interfaces	Not specified	LPDDR4/4X, SDIO	LPDDR4, QSPI, SRAM	LPDDR4, QSPI, SDIO/eMMc
Operating Temperature	-40C to +125C	Not specified	Not specified	Not specified
DSP Support	eFPGA core for adaptable operators	Vector DSP, 256 MACs @ 700MHz	Programmable HiFi 3 DSP	Quad-core ARM A53, Vector DSP, Computer Vision Processor, HiFi 3 DSP
IP Availability	TSMC IP Alliance Member	Not specified	Not specified	Not specified
Form Factor	Arrays from 1x1 to larger sizes	Not specified	3.1mm x 2.5mm 42-ball WLBGA, 5mm x 5mm 40-pin QFN	FCBGA 1369 balls; 31mmx31mm, 0.8mm pitch; 5mm x 5mm 40-pin QFN - AEC-Q100 Grade 3 qualified automotive SKU

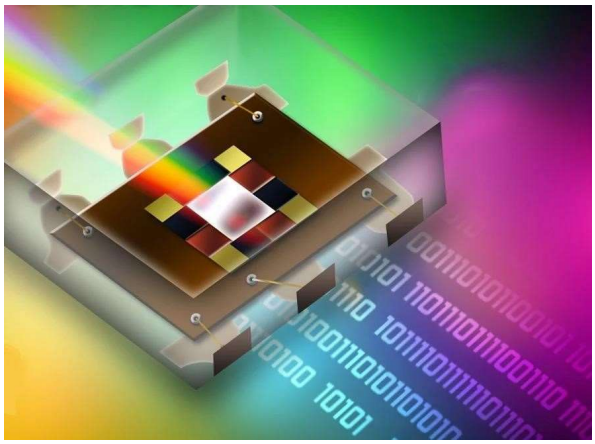
Conclusion:

In conclusion, the discussed NPUs, including InferX™ Vision AI, Hailo-15™ AI Vision Processor, Syntiant NDP120 Neural Decision Processor, and MLSoc AI, share common priorities such as real-time processing, power efficiency, and adaptability for evolving neural network architectures. InferX™ stands out with GPU-like performance and an eFPGA core for adaptability, while Hailo-15™ excels in a standard camera power envelope with versatility for various use cases. Syntiant NDP120 focuses on always-on applications and advanced audio processing, and MLSoc AI serves as a purpose-built MLSoC platform with flexibility for running concurrent DNN models. These NPUs collectively contribute to advancing AI capabilities, particularly in autonomous cars, by delivering high-performance, power-efficient, and adaptable neural network processing solutions with diverse implementations and specialized strengths.

OPTICAL INTERCONNECT

WORKING PRINCIPLE

Data is converted from electrical signals to optical signals through a modulator. Optical fibers typically have a much greater data-carrying capacity compared to copper, enabling the consolidation of multiple data streams onto a single optical fiber through various multiplexing techniques. At the receiving end, the optical signal undergoes demultiplexing before being directed to a photodetector, where the data is converted back to the electrical domain.



Optical interconnects are emerging as a prominent contender. While other alternatives, such as wireless network-on-chip and traditional transmission lines, present challenges in terms of signal decay, limited bandwidth, and scalability issues, optical interconnects have garnered attention for their potential to overcome these limitations. We spend most of our energy in communication not logic level.

Interconnect Challenges in electrical

- Variability and Scaling: Fabrication variability has increased with nanoscale wires.
- Capacitance and Resistance: Decreasing wire width increases resistance, requiring increased wire height, leading to higher capacitance.
- Signal Coupling: Undesired signal coupling between neighboring wires introduces noise.
- Mitigation Strategies: Repeater use, wire-aware designs, and passive/active shielding address signal coupling issues.

- Power Consumption Concerns: Interconnects consume a significant portion of microprocessor power, limiting multicore.

Optical Interconnects as a Solution:

- Rationale for Optical Interconnects: Metal interconnects are becoming insufficient for many-core systems.

- Optical Interconnect Challenges: Power consumption and integration challenges remain, but rapid progress is noted.

- Key points which differ optical interconnect from traditional electrical interconnect are following:-

Types of connection	Bandwidth (per lane)	Energy Efficiency	Latency	Power consumption
Electrical Interconnect	Upto 16Gb/s	250 fJ/bit	32 ns	45.3 mW
Optical interconnect	Upto 256Gb/s	10 fJ/bit	< 2 ns	11.5 mW

Benefits

- Enhanced Bandwidth Density: Optical interconnects, particularly Mode-Division Multiplexing (MDM), offer superior bandwidth density compared to traditional electrical interconnects. This results in increased data transfer rates within the chiplet architecture.

- Energy Efficiency: Optical interconnects contribute to improved energy efficiency by reducing power consumption, especially with the proposed multimode coupling solution based on a rectangular core few-mode fiber (RCF) and an integrated multimode coupler.

- Compact Footprint: Silicon photonics and the use of advanced materials enable the creation of compact, space-saving optical interconnects, facilitating higher integration levels within the chiplet architecture.

- Low Latency: Optical interconnects can provide lower latency compared to electrical interconnects, enhancing the overall speed and responsiveness of chip-to-chip communication.

- Scalability: The proposed integration plan is designed to be scalable, allowing for future upgrades and seamless integration of emerging optical technologies, ensuring the longevity and relevance of the chiplet architecture.

Challenges and Mitigations

Multimode Chip-Fiber Mismatch: The potential mismatch between transverse modes on the MDM chip and linear polarization modes in the circular core few-mode fiber (CCF) can pose a challenge. Mitigation: The proposed solution using a rectangular core few-mode fiber (RCF) addresses this mismatch, offering a promising approach for efficient multimode coupling.

Differential Group Delays (DGDs): High DGDs in conventional circular core few-mode fibers can complicate digital signal processing (DSP). Mitigation: The use of RCF with ultra-low DGDs reduces DSP computation complexity, contributing to energy-efficient MDM transmission.

Fabrication Complexity: The fabrication process of advanced optical components may introduce complexities. Mitigation: Continuous collaboration with industry partners and research institutions ensures access to optimized fabrication techniques.

Alignment Challenges: Achieving precise alignment for efficient multimode coupling can be challenging. Mitigation: Robust testing protocols, alignment automation tools, and technician training programs are implemented to address and minimize alignment issues during fabrication and deployment.

Wavelength Sensitivity: Sensitivity to wavelength variations, as observed in the TM10 mode channel, can impact the normalized spectrum. Mitigation: Utilizing optimized schemes, such as subwavelength-grating and meta-structures, can mitigate wavelength sensitivity, ensuring a more stable and broader operation bandwidth.

Optimizing Optical Interconnection Networks for Chip Multiprocessors

The contemporary trend of integrating an escalating number of processing cores into a single die underscores the critical importance of designing a robust and efficient communication infrastructure among them. Recent research endeavors have predominantly concentrated on the development of packet-switched Networks-on-Chip (NoC) designs tailored for both general-purpose Chip Multiprocessors (CMP) and application-specific Systems-on-Chip (SoC). This research aims to optimize critical aspects such as NoC bandwidth and latency, which directly impact the overall application performance of these integrated systems.

Challenges in NoC Power Dissipation

However, the packaging constraints and the imperative to adhere to strict on-chip temperature limits pose significant challenges for future CMPs. Recent prototypes of CMPs with a substantial number of cores reveal that over 25% of the overall power is dissipated by the NoC. Moreover, the power dissipation of current NoC implementations, using conventional circuit techniques, is estimated to be excessively high (by a factor of 10) to meet the anticipated demands of future CMPs. Consequently, careful distribution of the limited on-chip power budget between computation and communication activities becomes paramount for enhanced performance-per-watt metrics.

Hybrid Photonic-Electronic NoC Design: Building upon these advantages, a novel approach proposes a hybrid NoC design. This design integrates a high-bandwidth circuit-switched photonic network with a low-bandwidth packet-switched electronic network. In this hybrid system, the electronic network handles small-size control and data packets, while the photonic network efficiently transfers large-size data messages between pairs of cores.

Operational Workflow of Photonic NoC: The operational workflow of the photonic NoC involves a systematic process:

1. Path Setup: Photonic circuits are reserved through the exchange of path-setup packets over the electronic network, followed by a short acknowledgment pulse over the photonic network.

2. Data Transfer: Large data transfers occur on the photonic circuit, reaching up to 960Gbps of transmission line rate per core through timedivision and wavelength-division multiplexing.

3. Path Teardown: The source releases the photonic circuit through the transmission of a tear-down packet. NoC Organization and Components:

n a 16-core CMP configuration, the NoC is organized with network interfaces (gateways) represented by black circles and photonic switches composed of Photonic Switching Elements (PSE) and Electronic Routers (ER). PSEs, leveraging silicon micro-ring resonators, facilitate light deflection/passing, while ERs manage electronic packets and PSE polarization during setup and teardown.

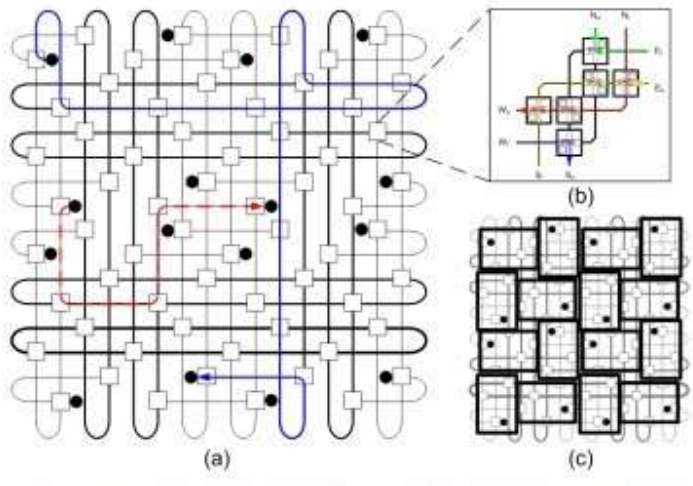


Figure 82 16 core Blocking Mesh NoC. (a) shortest (longest) path is marked dashed (solid) (b) basic non-blocking switch (c) core layout over the NoC

Overcoming Blocking Topology:

The proposed Blocking Mesh topology initially exhibits a blocking nature, limiting connectivity and potentially causing delays. However, the paper suggests mitigating this limitation by over-provisioning the network. Doubling the number of rows and columns improves NoC performance, reducing the blocking probability and enhancing overall system throughput.

Integration Challenges and Future Directions:

Despite the potential of photonic communication, the integration of optical devices on a chip presents challenges. Recent breakthroughs in CMOS-compatible silicon photonics, particularly with innovative small-footprint devices like silicon micro-ring resonators, show promise. The proposed layout involves considering 3D integration for optimized fabrication. Future work explores alternative non-blocking topologies and conducts a comparative analysis of performance and power efficiency. The paper concludes with a case study demonstrating the potential benefits of using silicon photonics, analyzing the total execution time and power dissipation for a hypothetical 36-core CMP in a future 22nm technology process.

Core Model for Photonic NoCs

The core model is designed for photonic Network-on-Chip (NoC) systems, aimed at facilitating high-bandwidth, low-latency communication channels for extensive data transfers between processor cores. Each core in this model is a multi-threaded processor, capable of parallel thread execution, with each thread independently initiating data transfers to other cores.

Traffic Generator:

- Simulates core thread behavior, generating data transfer requests during processing.
- The number of threads per core is a simulation parameter.
- Each thread can request one connection at a time, preventing the simultaneous request from exceeding the core's thread count.
- Communication requests follow a Poisson process with uniformly-distributed destinations.
- Message length can be fixed or randomly set with an exponential probability distribution.
- Requests are stored in a finite-size back-pressuring FIFO queue.

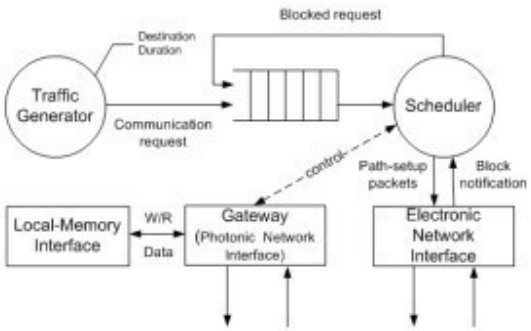


Figure 83 Model of Multi-threaded processing cores

Scheduler:

- Extracts requests from the FIFO to generate relative path-setup packets.
- Attempts to inject/eject packets into/from the network via the Electronic Network Interface.
- Blocked requests are re-enqueued into the FIFO.
- Aims to prevent head-of-line (HoL) blocking, a known issue in switching networks.
- Monitors a window of packets at the head of the FIFO to establish connections based on arrival time.
- After successful communication, the process restarts from the oldest packet to avoid HoL blocking.

Gateway (Photonic Network Interface):

- Represents the Photonic Network Interface, responsible for sending/receiving photonic messages to/from the NoC.
- Reads/writes data from/into the Local-Memory Interface.

Workflow

Traffic Generation:

- Core threads generate communication requests based on a Poisson process.
- Requests are placed in the back-pressuring FIFO queue.

Scheduling:

- The Scheduler processes requests from the FIFO to create path-setup packets.

- Attempts to inject/eject packets into/from the network, avoiding HoL blocking.
- Blocked requests are re-enqueued.

Gateway Operation:

- The Photonic Network Interface (Gateway) manages the actual transmission/reception of photonic messages to/from the NoC.
- Data is read/written through the Local-Memory Interface.

Non-Blocking Topologies for Photonic NoC

Crossbar

- Utilizes an 8×8 matrix of switches for a 16-core Chip Multiprocessor (CMP).
- Strictly non-blocking with O(N^2) complexity.
- Limited scalability in terms of resources and maximum/average path length.
- Issues include electronic connection lengths impacting power dissipation.

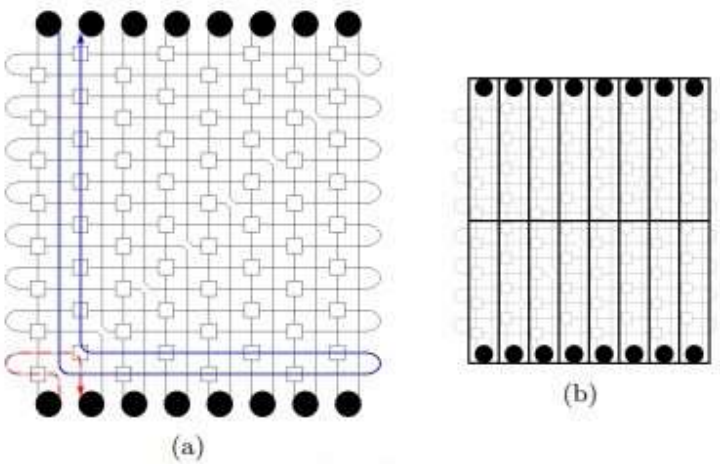


Figure 84 (a) 16- core Crossbar (b) core layout over the NoC

Non-Blocking Mesh

- Constructed using non-blocking switches and simplified injection/ejection policies.
- Achieves non-blocking topology with only two cores injecting and two cores ejecting per row and column.
- For a 16-core CMP, it forms an 18×18 mesh of switches with 36 gateway switches.
- Offers improved scalability compared to Crossbar, reducing average path length between cores.

Number of Cores	Blocking Mesh	Crossbar	Non-blocking Mesh
16	144 switches 8 avg path	64 switches 12 avg path	80 switches 6 avg path
36	324 switches 12 avg path	324 switches 27 avg path	360 switches 11 avg path
64	576 switches 16 avg path	1024 switches 48 avg path	1088 switches 18 avg path

- The Blocking Mesh exhibits superior scalability with the number of switches growing linearly and the path length scaling as √N.
- Non-Blocking Mesh, while maintaining O(N^2) complexity, offers a substantial improvement over the Crossbar by reducing the average path length.
- However, as the network size increases, the Non-Blocking Mesh faces challenges with signal integrity and resource requirements, impacting path setup times.

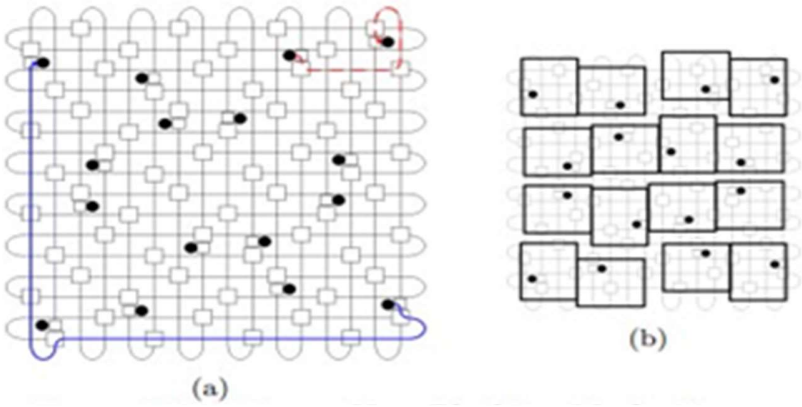


Figure 85 (a) 16 core Non-Blocking Mesh (b) core layout over the NoC

Comparative Performance Analysis

Experimental Setup:

In a study involving 36 cores engaged in fixed-size DMA transfers (16kBytes) at a line rate of 960Gbps, resulting in a photonic message duration of 134ns, four distinct scenarios were investigated: Blocking Mesh (single-thread), Blocking Mesh (multi-threaded), Cross

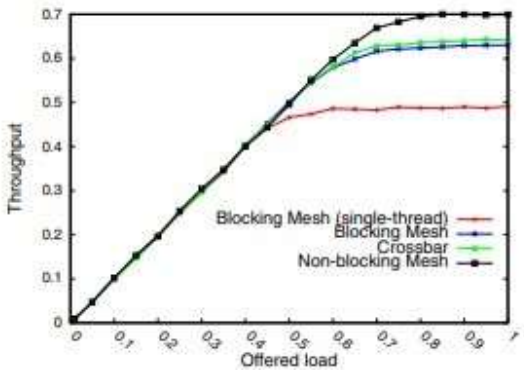


Figure 86 Throughput per core of various 36-core NoC topologies

Key Findings:

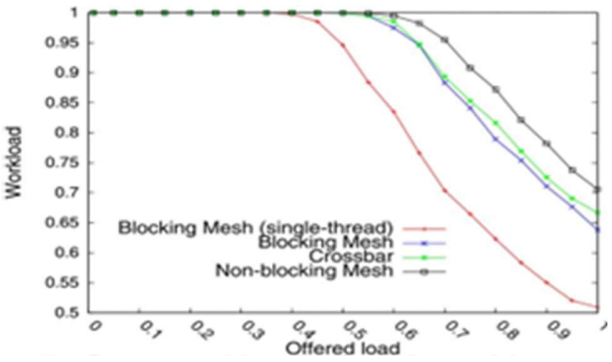


Figure 87 Core workload of various 36-core NoC topologies

- **Multithreaded Cores Impact:** Multithreaded cores demonstrated a notable enhancement in the utilization of photonic NoC bandwidth, resulting in a substantial 26% increase in throughput-per-core. This improvement is attributed to the ability of multithreaded cores to better exploit the high bandwidth offered by a photonic NoC.
- **Topology Performance**: Crossbar and Blocking Mesh exhibited comparable performance, challenging the conventional expectation that non-blocking topologies inherently outperform others. Non-Blocking Mesh, surprisingly, achieved a 13% throughput gain due to shorter gateway distances, mitigating some of the drawbacks associated with non-blocking topologies.
- **Workload Bottleneck**: The workload of a core, reflecting its computational efficiency, became a bottleneck when the offered load reached a threshold, approximately 0.6 in this case. Beyond this threshold, the network struggled to handle the increased traffic load efficiently.

FFT Computation

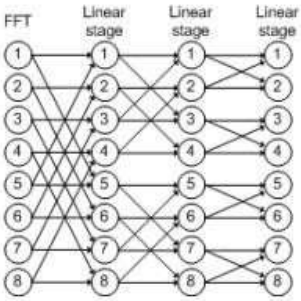


Figure 8. Butterfly scheme of Cooley-Tukey's algorithm for $M = 8$.

- **FFT Algorithm Overview:** The Cooley-Tukey FFT algorithm processes an array of input samples, dividing the task among cores. The algorithm consists of a phase where each core handles a portion of the samples, followed by iterations involving computation and communication steps, exchanging data in a butterfly scheme.
- **Experimental Setup:** Assuming a hypothetical CMP in a future 22nm technology process with 3D Integration, each core is estimated to have a local memory of about 0.5GBytes. Scaling assumptions suggest integrating 36 cores comparable to the first generation of the IBM Cell multi-core processor.
- **Performance Estimations:** Using Bailey's FFT algorithm within each core, the total execution time for a 229 double-precision sample FFT in the Non-Blocking Mesh is estimated at 66ms, with 14ms dedicated to butterfly data exchanges.

- **Comparison with Blocking Mesh**: Simulations reveal that a CMP equipped with a Blocking Mesh takes 74.6ms to complete the FFT computation, indicating an 8.6ms increment for butterfly data exchanges compared to the Non-Blocking Mesh. This difference is attributed to conflicts within the network in the blocking topology.
- **Photonic Transmission Line Rate Considerations**: For photonic transmission at 960Gbps, the power consumption is estimated based on available transceiver technology. Three scenarios with energy consumption values of 0.8pJ/bit, 0.4pJ/bit, and 0.2pJ/bit are considered.
- **Power Consumption and Line Rate Analysis**: For a Non-Blocking Mesh, transferring 32 blocks of 256MBytes at 960Gbps requires less than 24.5W, translating to about 770mW per connection. The study compares FFT computation time ratios and power ratios for different line rates, highlighting the efficiency of photonic implementations. (Association & Ann Arbor) (Warnock J. Circuit design challenges at the 14 nm technology node; Proceedings of the 48th Design Automation Conference; San Diego, CA, USA. 5 June 2011.) (Ho R., Mai K., Horowitz M. Efficient on-chip global interconnects; Proceedings of the 2003 Symposium on VLSI Circuits. Digest of Technical Papers; Kyoto, Japan. 12–14 June 2003; pp. 271–274)

Comparative Analysis of Photonic and Electronic NoC Performance

After evaluating the performance of a Chip Multiprocessor (CMP) with a photonic network, this analysis explores the idea of replacing it with an equivalent electronic network. The study focuses on the Fast Fourier Transform (FFT) computation, considering channel utilization and communication efficiency in both photonic and electronic scenarios.

- **Performance of Electronic NoCs:** Due to the persistent channel utilization in FFT sub-array transfers, a circuit-switched data network outperforms a packet-switched NoC. The equivalent electronic network replaces photonic components with functionally similar electronic ones. The analysis accounts for delay and

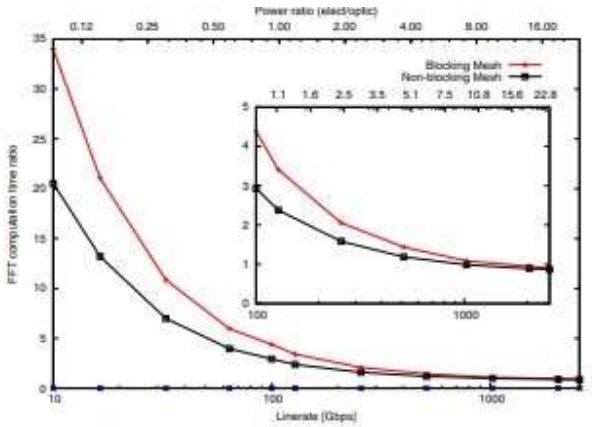


Figure 9. FFT-computation- time ratio and power ratio as a function of line rate for the electronic implementation of two 36-core topologies (blocking and non-blocking) with respect to an equivalent photonic Non-Blocking Mesh, which takes $66ms$ and dissipates $24.5W$ to complete the same task with a $960Gbps$ line rate.

Figure 88 FFT computation time ratio and power ratio as a function of line rate for the electronic implementation of two 36 core topologies (blocking and non-blocking) with respect to an equivalent photonic non-blocking Mesh, which takes 66ms and dissipates 24.5W to complete the same task with a 960 Gbps line rate.

power consumption, assuming an ideal electronic implementation without delay or power consumption.

- **Communication Characteristics:**
  Considering the length of optimally repeated wires in a 22nm technology and assuming an energy consumption of 0.25pJ/bit/mm, communication over the equivalent electronic circuit switched NoC exhibits millisecond-level message durations. The computation time depends solely on the line rate for both photonic and electronic networks.

Photonic power $[pJ/bit]$	Power efficiency gain	Performance gain
0.8	10×	3×
0.4	20×	4.5×
0.2	40×	8.5×

Figure 89 Performance and power gains as function of the photonic power consumption projections

- **Performance Comparison:**
  Fig. illustrates the difference in computing a 229-sample FFT with an electronic NoC, varying the core's line rate for different topologies. The y-axis represents the ratio of total execution time to the reference time of 66ms for the photonic Non-Blocking Mesh at 960Gbps. The chart also includes the power dissipation ratio for the electronic NoC.

- **Gain in Performance-Per-Watt:**
  - To assess the performance-per-watt gain, two scenarios are considered:
  - Achieving the same execution time as the photonic NoC results in an electronic NoC dissipating 10 times more power (7.6W per connection), exceeding the total power budget for the CMP.
  - Achieving the same power dissipation as the photonic NoC requires the electronic NoC to operate at a reduced line rate of 100Gbps, taking about 190ms to complete the FFT computation—three times more than the reference network.

- **Power and Performance Scaling:**
  - Table presents the scaling of power and performance gains with different projections of future photonic transceiver power consumption, showing the potential for substantial improvements.

Conclusions

On-chip photonic communication emerges as a promising solution for meeting the communication demands of future high-performance Chip Multiprocessors (CMPs). Through our simulation-based assessment, we determine that a photonic Network-on-Chip (NoC) is particularly well-suited for connecting a constrained number of complex multi-threaded cores, especially when the core count is limited, such as not exceeding 36 cores. In this context, a photonic NoC with a non-blocking topology outperforms a blocking topology, providing superior performance without additional design complexities. Importantly,

regardless of the chosen topology, our findings indicate that for communication-intensive applications on future large CMPs, a photonic NoC serves as a viable alternative to electronic NoC, seamlessly delivering high-bandwidth and low-power connectivity with off-chip devices.

SIMULATION

How data transfer through optical Interconnect

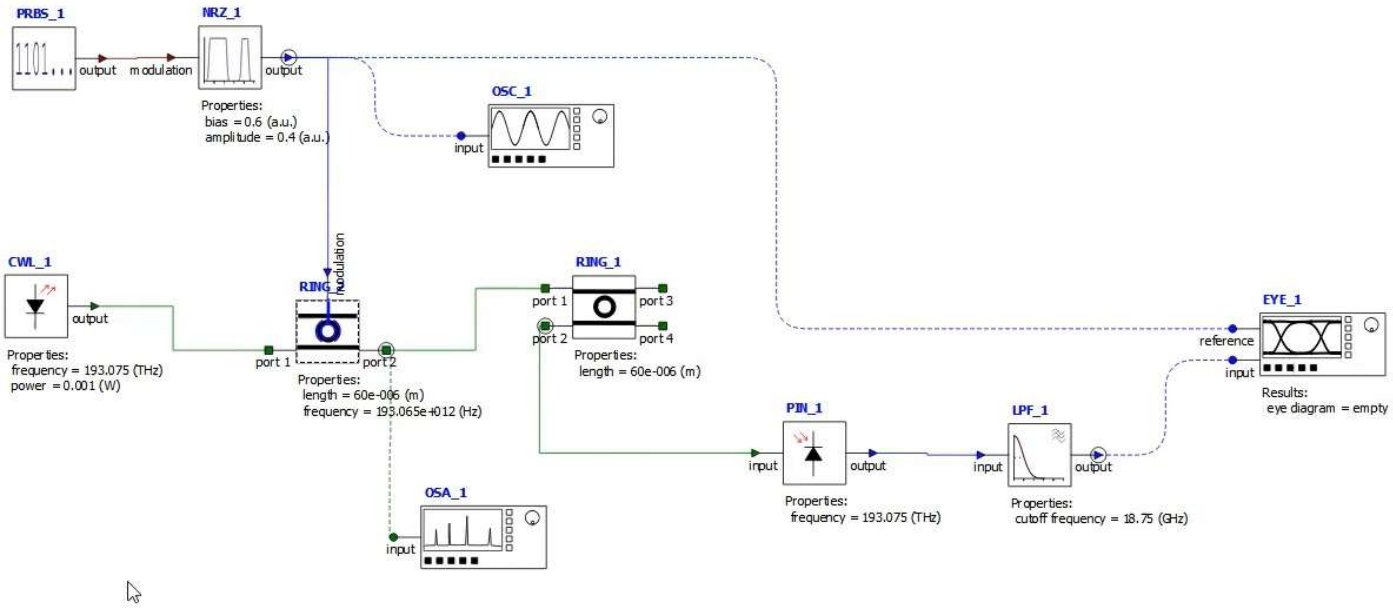


Figure 90 Schematic diagram of optical interconnect

How data covert from digital to optical signal

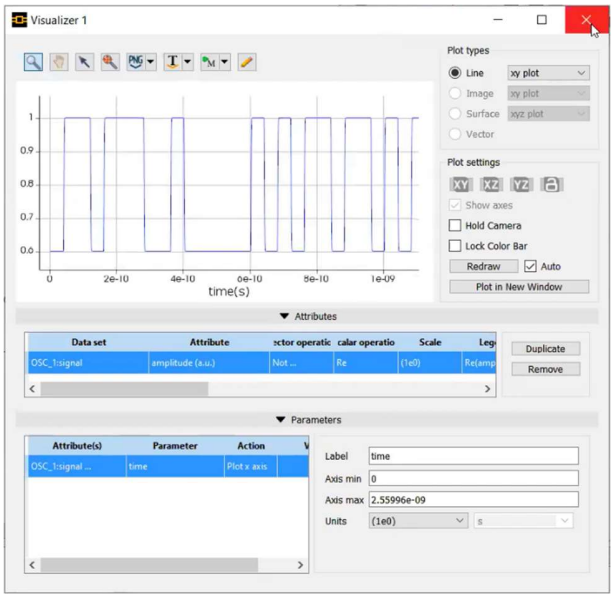


Figure 92 Random Generated Digital Signal

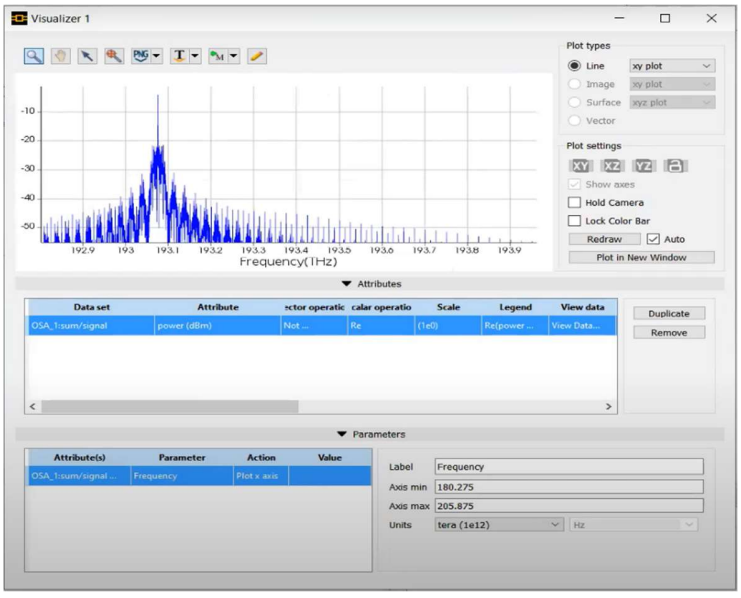


Figure 91 Optical signal generated by Ring Modulator

SIMULATION 1

A bitrate of 2.5e+10 bits/s, 1024 samples per bit, and a sequence length of 64 yielded a perfect graph with no distortion. This scenario demonstrated the system's capability to generate a high-quality signal suitable for seamless conversion into digital format.

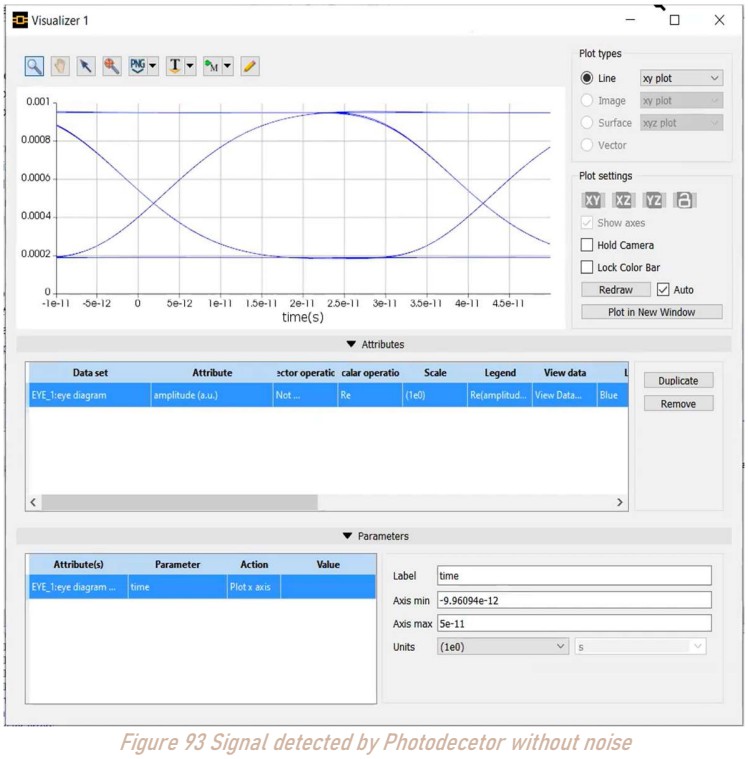


Figure 93 Signal detected by Photodecetor without noise

SIMULTION 2

In **Simulation 2 we** introduced noise into the system while maintaining the same parameters as Simulation 1. Although some distortion was observed in the output graph, the signal remained convertible into digital form with the caveat of requiring additional precision. This highlighted the system's resilience to noise and the potential for optimization to enhance precision

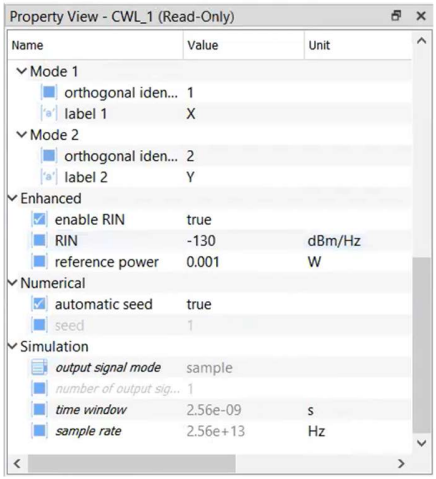


Figure 94 When noise is introduced

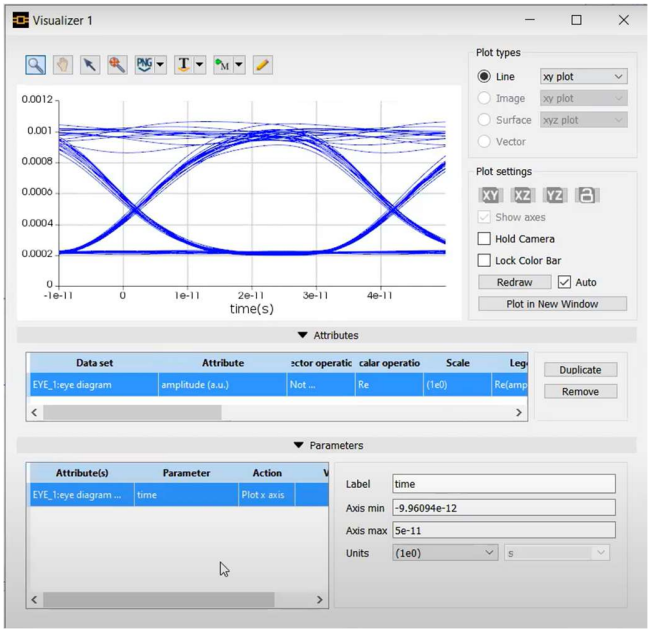


Figure 95 Signal detected by Photodector with Noise

SIMULATION 3

In **Simulation 3**, with an increased bitrate of 6.0e+10 bits/s, exhibited very great distortion, rendering the signal impractical for direct digital conversion. The distortion observed in this scenario emphasized the importance of carefully choosing parameters to avoid signal degradation.

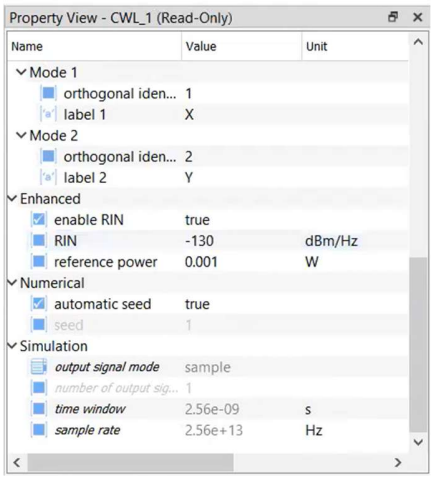


Figure 97 Noise plus increased bit rate

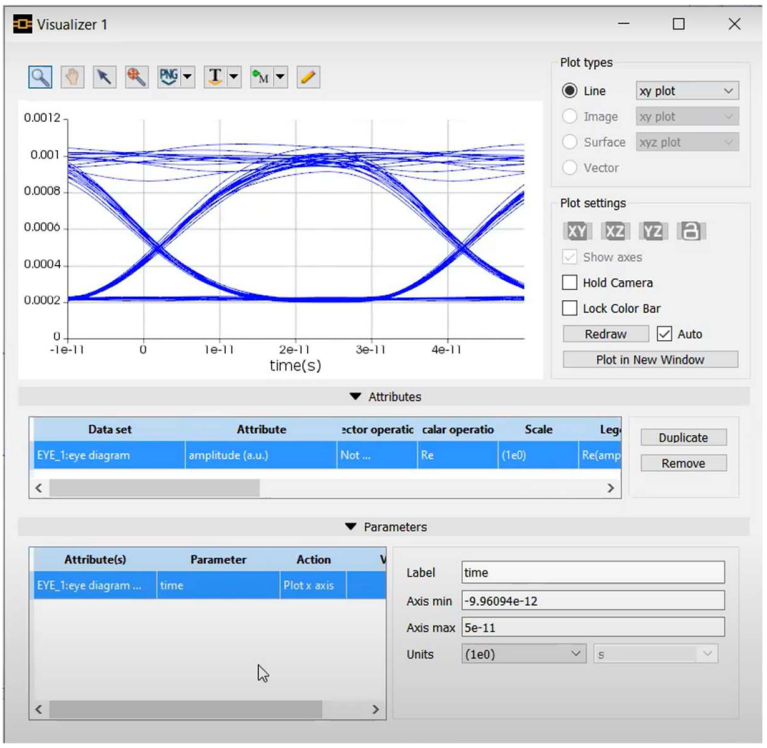


Figure 96 Noise plus increased bit rate signal detected by photodetector

Simulation Result	Bitrate (bits/s)	Samples per Bit	Sequence Length	Type Digital Signal	RIN (Noise)	Temperature (K)	Nature of Graph	Convertible into Digital
1	2.5e+10	1024	64	Random Generated Bit	No	300	Perfect Graph - No Distortion	Yes
2	2.5e+10	1024	64	Random Generated Bit	-130 dbm/Hz	300	Some Distortion Observed	Yes (Precision Required)
3	6.0e+10	1024	64	Random Generated Bit	-130 dbm/Hz	300	Severe Distortion Noted	No (Signal Too Distorted)

How data transfer from different data line to another

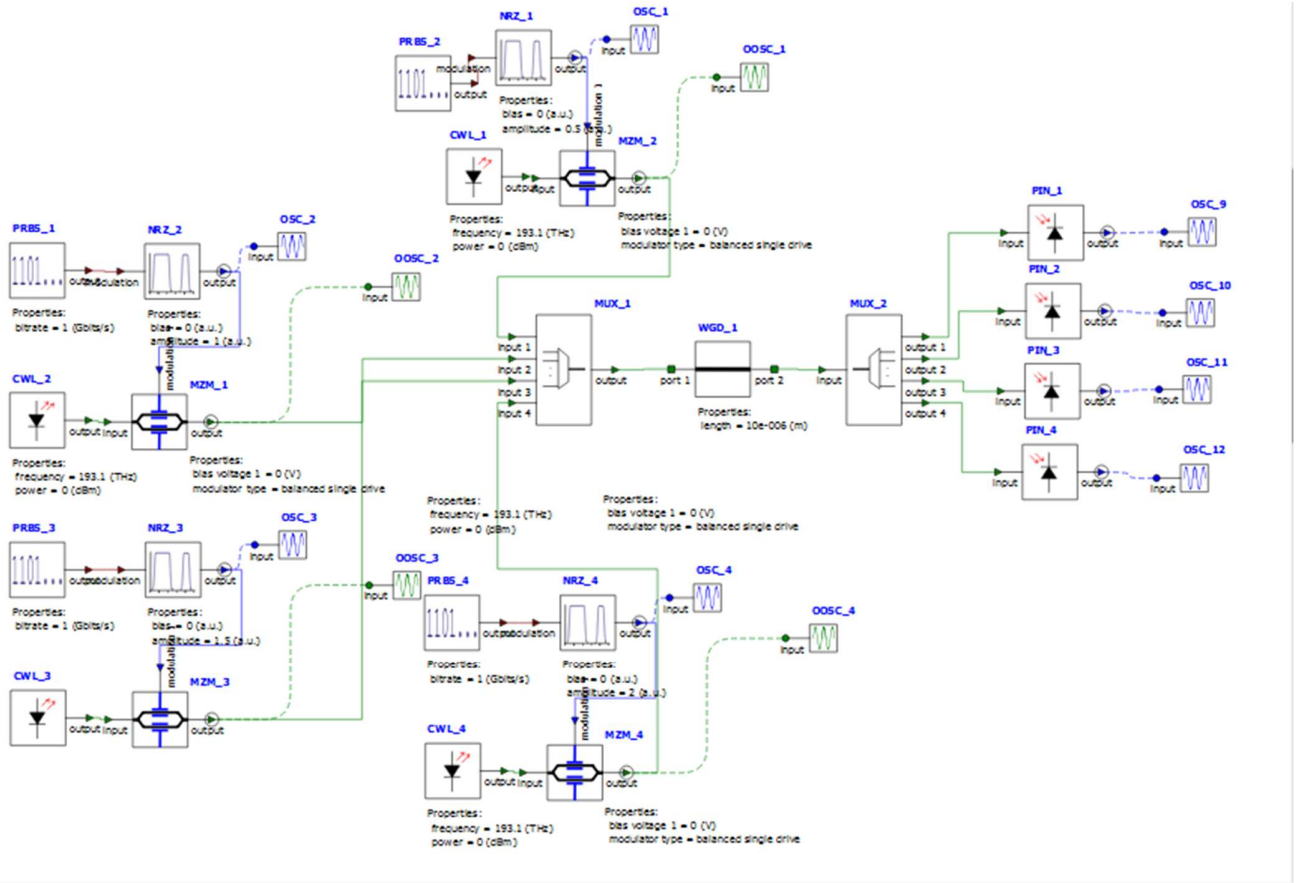


Figure 98 Schematic diagram of optical interconnect simulation

CONCLUSION

The conducted simulations, considering varying bitrates, noise, and signal characteristics, reveal that optical interconnect is feasible under specific conditions. Optimally configured systems demonstrate robustness to noise, enabling reliable optical-to-digital conversion. However, careful parameter selection is crucial to avoid significant distortion, emphasizing the need for a balanced approach in designing optical interconnects for effective high-speed data transmission

BIBLIOGRAPHY

G. E. Moore et al., “Cramming more components onto integrated circuits,” 1965. . (n.d.).

S. Mirabbasi, L. C. Fujino, and K. C. Smith, “Through the looking glass—the 2022 edition: Trends in solid-state circuits from ISSCC,” IEEE Solid-State Circuits Magazine. (n.d.).

[1] T. Li, J. Hou, J. Yan, R. Liu, H. Yang, and Z. Sun, “Chiplet heterogeneous integration technology—status and challenges,” Electronics, vol. 9, no. 4, p. 670, 2020. . (n.d.).

(n.d.). *[22] A. H. A. Aboushady et al., "High-Performance Parallel Interfaces for Chiplet-Based Systems," in IEEE Transactions on Very Large Scale Integration .*

“Bunch of wires (bow) phy specification,” https://opencomputeproject. github.io/ODSA-BoW/bow specification.html. . (n.d.).

(n.d.). *• [22] A. H. A. Aboushady et al., "High-Performance Parallel Interfaces for Chiplet-Based Systems," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 27, no. 11, pp. 2975-2988, Nov. 2019. doi: 10.1109/TVLSI.2019.2914269 .*

20200910_Automotive Networking Technologies. (n.d.). Retrieved from https://cdn.intrepidcs.net/events/Webinars/Automotive_Networking_Technologies_20200910.pdf

4 Ways Augmented Reality Will Revolutionise the Automotive Industry. (n.d.). Retrieved from https://www.vanarama.com/blog/cars/4-ways-augmented-reality-will-revolutionise-the-automotive-industry

A Perspective on Ethernet in Automotive Communications—Current Status and Future Trends. (n.d.). Retrieved from https://www.mdpi.com/2076-3417/13/3/1278#:~:text=In%20light%20of%20this%20context%2C%20Automotive%20Ethernet%20is,rates%2C%20from%2010%20Mbps%20up%20to%2010%20Gbps

Ablassmeier, M., Poitschke, T., Wallhoff, F., Bengler, K., & Rigoll, G. (n.d.). *Eye Gaze Studies Comparing Head-Up and Head-Down Displays in Vehicles*. Retrieved from https://ieeexplore.ieee.org/document/4285134

AMD ON WHY CHIPLETS—AND WHY NOW. (n.d.). Retrieved from https://www.nextplatform.com/2021/06/09/amd-on-why-chiplets-and-why-now/

AMD Unveils Speedier Chiplet Design With High-Bandwidth Interconnects. (n.d.). Retrieved from https://www.enterpriseai.news/2021/01/04/amd-unveils-speedier-chiplet-design/

An Edge-Computing Based Architecture for Mobile Augmented Reality. (n.d.). Retrieved from https://ieeexplore.ieee.org/document/8612452

An in-depth comparison of LiDAR, Cameras, and Radars' technology. (n.d.). Retrieved from https://www.outsight.ai/insights/how-does-lidar-compares-to-cameras-and-radars

Architecture and key terminal technologies of 5G-based internet of vehicles. (n.d.). Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0045790621003918

Architecture for High-Throughput Low-Latency Big Data Pipeline on Cloud. (n.d.). Retrieved from https://www.apexon.com/blog/architecture-for-high-throughput-low-latency-big-data-pipeline-on-cloud/

ARVE: Augmented Reality Applications in Vehicle to Edge. (n.d.). Retrieved from https://core.ac.uk/download/pdf/245130002.pdf

Association, T. N., & Ann Arbor, M. U. (n.d.). Retrieved from https://scholar.google.com/scholar_lookup?title=The+National+Technology+Roadmap+for+Semiconductors&publication_year=1997&

Augmented Reality Coming to a Windshield Near You. (n.d.). Retrieved from https://spectrum.ieee.org/augmented-reality-car-hud?utm_campaign=post-teaser&utm_content=wvcltaqo

AUTONOMOUS VEHICLES AND THE CHALLENGES IN INDIA. (n.d.). Retrieved from https://lexplosion.in/autonomous-vehicles-and-the-challenges-in-india/

Autonomous vehicles on Indian roads still a distant dream. (n.d.). Retrieved from https://www.financialexpress.com/business/express-mobility-autonomous-vehicles-on-indian-roads-still-a-distant-dream-3125903/

Bulut, M. (n.d.). *A Review of Vapor Chambers.* Retrieved from https://www.researchgate.net/publication/325834311_A_Review_of_Vapor_Chambers

Bunch of Cruise cars stuck on Gough by Robin. (n.d.). Retrieved from https://www.reddit.com/r/sanfrancisco/comments/vnmpf1/bunch_of_cruise_cars_stuck_on_gough_by_robin/

Camera, Radar and LiDAR: A Comparison of the Three Types of Sensors and Their Limitations. (n.d.). Retrieved from https://autocrypt.io/camera-radar-lidar-comparison-three-types-of-sensors/

Chiplet Technology & Heterogenous Integration. (n.d.). Retrieved from https://nepp.nasa.gov/docs/etw/2021/15-JUN-21_Tues/1500_Ramamurthy-Chiplet-Technology-v3.pdf

Comer, D. (n.d.). *Essentials of Computer Architecture 2nd Edition*. Retrieved from https://www.amazon.com/Essentials-Computer-Architecture-Douglas-Comer/dp/1138626597

Comparison of PIM architectures based on Chiplet. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9609218/table/micromachines-13-01790-t001/?report=objectonly

Designing C-V2X Communication Systems: Key Engineering Considerations and Best Practices. (n.d.). Retrieved from https://www.wevolver.com/article/designing-c-v2x-communication-systems-key-engineering-considerations-and-best-practices

Dia-torus: A novel topology for network on chip design. (n.d.). Retrieved from https://www.researchgate.net/publication/303870995_Dia-torus_A_novel_topology_for_network_on_chip_design#full-text

Difference between LIN, CAN and FlexRay Protocols. (n.d.). Retrieved from https://prodigytechno.com/difference-between-lin-can-and-flexray-protocols/

Display and Interaction Systems Bosch. (n.d.). Retrieved from https://www.bosch-mobility.com/en/solutions/infotainment/display-and-interaction-systems/

Driverless Car Market Leaders Innovating The Transportation Industry. (n.d.). Retrieved from https://www.forbes.com/sites/cindygordon/2021/12/29/driverless-car-market-leaders-innovating-the-transportation-industry/?sh=242bfdbe137f

Edge Assisted Mobile Semantic Visual SLAM. (n.d.). Retrieved from https://ieeexplore.ieee.org/document/9155438

Enabling interposer-based disintegration of multi-core processors. (n.d.). Retrieved from https://ieeexplore.ieee.org/document/7856626

Everything You Need to Know About In-Vehicle Infotainment Systems. (n.d.). Retrieved from https://www.einfochips.com/blog/everything-you-need-to-know-about-in-vehicle-infotainment-system/

FPGA implementation of I2C & SPI protocols: A comparative study. (n.d.). Retrieved from https://ieeexplore.ieee.org/document/5410881

Getting ready for next-generation E/E architecture with zonal compute. (n.d.). Retrieved from https://www.mckinsey.com/industries/semiconductors/our-insights/getting-ready-for-next-generation-ee-architecture-with-zonal-compute

GM's Cruise Falls Down During Concert Cell Overload. (n.d.). Retrieved from https://www.forbes.com/sites/bradtempleton/2023/08/14/gms-cruise-falls-down-during-concert-cell-overload-heres-how-to-fix-it/?sh=30c3c966239c

György Bognár, G. T. (n.d.). *A Novel Approach for Cooling Chiplets .* Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10192284

Hannah, B. (n.d.). *Evaluating Head-Up Displays across Windshield Locations*. Retrieved from https://par.nsf.gov/servlets/purl/10184743

Heterogeneous Die-to-Die Interfaces: Enabling More Flexible Chiplet Interconnection Systems. (n.d.). Retrieved from https://dl.acm.org/doi/fullHtml/10.1145/3613424.3614310

Ho R., Ken M., Horowitz M. Managing wire scaling: A circuit perspective; Proceedings of the IEEE 2003 International Interconnect Technology Conference; Burlingame, CA, USA. 4 June 2003; pp. 177–179. (n.d.). Retrieved from https://scholar.google.com/scholar_lookup?journal=Proceedings+of+the+IEEE+2003+International+Interconnect+Technology+Conference&title=Managing+wire+scaling:+A+circuit+perspective&author=R.+Ho&author=M.+Ken&author=M.+Horowitz&pages=177-179&

Ho R., Mai K., Horowitz M. Efficient on-chip global interconnects; Proceedings of the 2003 Symposium on VLSI Circuits. Digest of Technical Papers; Kyoto, Japan. 12–14 June 2003; pp. 271–274. (n.d.). Retrieved from https://scholar.google.com/scholar_lookup?journal=Proceedings+of+the+2003+Symposium+on+VLSI+Circuits.+Digest+of+Technical+Papers&title=Efficient+on-chip+global+interconnects&author=R.+Ho&author=K.+Mai&author=M.+Horowitz&pages=271-274&

Ho R., Mai K.W., Horowitz M.A. Proceedings of the IEEE. Volume 89. IEEE; Piscataway, NJ, USA: 2001. The future of wires; pp. . (n.d.). Retrieved from https://scholar.google.com/scholar_lookup?title=Proceedings+of+the+IEEE&author=R.+Ho&author=K.W.+Mai&author=M.A.+Horowitz&publication_year=2001&

Instruction Latency and How It Affects CPU Performance. (n.d.). Retrieved from https://itigic.com/instruction-latency-and-how-it-affects-cpu-performance/

Introduction to Cellular V2X. (n.d.). Retrieved from https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/c-v2x_intro.pdf

Ismail Y.I., Friedman E.G., Neves J.L. Dynamic and short-circuit power of CMOS gates driving lossless transmission lines. . (n.d.). Retrieved from https://scholar.google.com/scholar_lookup?journal=Proceedings+of+the+8th+Great+Lakes+Symposium+on+VLSI&title=Dynamic+and+short-circuit+power+of+CMOS+gates+driving+lossless+transmission+lines.+IEEE+Transactions+on+Circuits+and+Systems+I:+Fundamental+Theory

Jacinto 7 Camera Capture and Imaging Subsystem. (n.d.). Retrieved from https://www.ti.com/lit/an/spracx9/spracx9.pdf?ts=1702362244630&ref_url=https%253A%252F%252Fwww.ti.com%252Ftool%252FSK-TDA4VM

John D.O. Research Challenges for On-Chip Interconnection Networks. 2007. [(accessed on 10 December 2021)]. pp. 96–108. (n.d.). Retrieved from https://www.computer.org/csdl/magazine/mi/2007/05/mmi2007050096/13rRUxAASPi

Kite: A Family of Heterogeneous Interposer Topologies Enabled via Accurate Interconnect Modeling. (n.d.). Retrieved from https://ieeexplore.ieee.org/document/9218539

Larimi, S. S. (n.d.). *Understanding Power Consumption and Reliability of High-Bandwidth Memory with Voltage Underscaling.* Retrieved from https://www.researchgate.net/publication/354108335_Understanding_Power_Consumption_and_Reliability_of_High-Bandwidth_Memory_with_Voltage_Underscaling

Modern GPU Architecture. (n.d.). Retrieved from https://download.nvidia.com/developer/cuda/seminar/TDCI_Arch.pdf

Performance Comparison of AMBA Bus-Based System-On-Chip Communication Protocol. (n.d.). Retrieved from https://ieeexplore.ieee.org/document/5966487

Pioneering Chiplet Technology. (n.d.). Retrieved from https://ieeexplore.ieee.org/document/

Qualcomm SA6155P. (n.d.). Retrieved from https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/qul7413_sa6155_productbrief_r2.pdf

S. Naffziger, N. Beck, T. Burd, K. Lepak, G. H. Loh, M. Subramony, and S. White, "Pioneering chiplet technology and design for the AMD EPYC™ and RYZEN™ processor families: Industrial product," in 2021 ACM/IEEE 48th Annual International Symposium on Co. (n.d.).

Semiconductor Design & Verification Articles. (n.d.). Retrieved from https://www.intrinsix.com/blog/chiplet-interfaces

Smith, M. (n.d.). *Head-Up vs. Head-Down Displays: Examining Traditional Methods of Display Assessment While Driving*. Retrieved from https://dl.acm.org/doi/abs/10.1145/3003715.3005419

Synopsys XSR PHY IP. (n.d.). Retrieved from https://www.synopsys.com/dw/ipdir.php?ds=dwc_usr-xsr_phy

TAP-2.5D: A Thermally-Aware Chiplet Placement Methodology for 2.5D Systems. (n.d.). Retrieved from https://www.researchgate.net/publication/347513282_TAP-25D_A_Thermally-Aware_Chiplet_Placement_Methodology_for_25D_Systems

The Shift Towards Zonal Network Architecture (Intrepid Tech Day '23). (n.d.). Retrieved from https://www.youtube.com/watch?v=vpmHJo8tUgE

Thermal Modeling of a Chiplet-Based Packaging With a 2.5-D Through-Silicon Via Interposer. (n.d.). Retrieved from https://ieeexplore.ieee.org/document/9785787

Warnock J. Circuit design challenges at the 14 nm technology node; Proceedings of the 48th Design Automation Conference; San Diego, CA, USA. 5 June 2011. (n.d.). Retrieved from https://scholar.google.com/scholar_lookup?journal=Proceedings+of+the+48th+Design+Automation+Conference&title=Circuit+design+challenges+at+the+14+nm+technology+node&author=J.+Warnock&

Waymo keeps 5G way in the background. (n.d.). Retrieved from https://www.lightreading.com/iot/waymo-keeps-5g-way-in-the-background

What is PCIe 4.0? PCI Express 4 explained. (n.d.). Retrieved from https://www.rambus.com/blogs/pci-express-4/

What is SerDes (Serializer/Deserializer)? (n.d.). Retrieved from https://www.synopsys.com/glossary/what-is-serdes.html

Which Data Interfaces Are Used for Chiplet Interconnects? (n.d.). Retrieved from https://resources.system-analysis.cadence.com/blog/which-data-interfaces-are-used-for-chiplet-interconnects

ZONAL ARCHITECTURE. (n.d.). Retrieved from https://f.hubspotusercontent30.net/hubfs/3474688/GuardKnox%20Zonal%20Architecture%20Whitepaper%20Executive%20Summary.pdf

Zone architectures power software-defined vehicles. (n.d.). Retrieved from https://www.electronicproducts.com/zone-architectures-power-software-defined-vehicles/