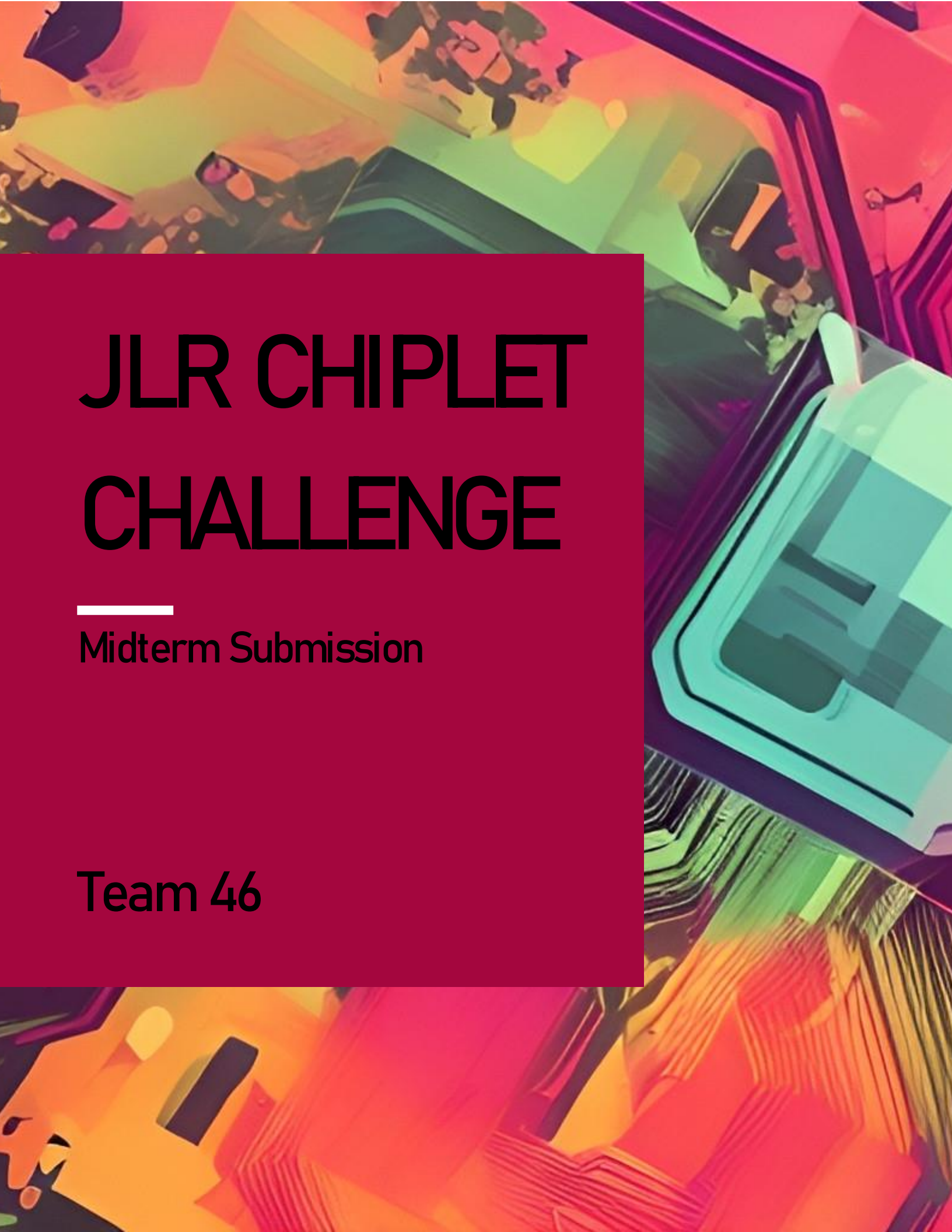


JLR CHIPLET CHALLENGE

Midterm Submission

Team 46

WHY CHIPLET TECHNOLOGY

"WE'RE BRINGING A SUPERCOMPUTER INTO THE CAR, BUT THIS IS WHERE THE CHALLENGE STARTS."

- BART PLACKLEY

Moore's Law had set the pace for the semiconductor industry for decades with a reliable generation-upon-generation increase in transistor density and a corresponding reduction in cost per transistor. But in recent times, the steady drum beat of Moore's Law has started to slow down. Whereas device density historically doubled every 18-24 months, the rate of recent silicon process advancements has declined. While improvements in device scaling continue, albeit at a reduced pace, the industry is simultaneously observing increases in manufacturing costs. In response, the industry is now seeing a trend toward reversing direction on the traditional march toward more integration. Instead, multiple industry and academic groups are advocating that systems on chips (SoCs) be "disintegrated" into multiple smaller "chiplets".

Over the last several years, AMD has been writing a new post-Moore's Law story that revises the historical trend of integrating more functionality per silicon chip and instead is disintegrating the traditional monolithic silicon chip into multiple smaller "chiplets."

The overall idea with chiplets is to take what would normally be a monolithic, single-die SoC, and then partition it into multiple smaller die or "chiplets" and then "reintegrate" them with some form of in-package interconnect to enable the collective to operate as a single, logical SoC.

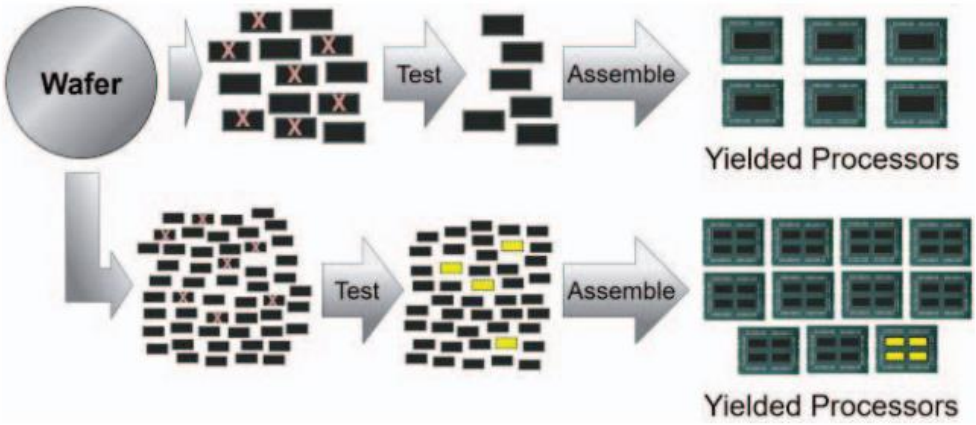


Fig. Illustrative construction of processors using (top) monolithic die and (bottom) reassembled chiplets. (The light/yellow chiplets represent chiplets that can run at higher clock speeds)

If an SoC with T transistors can be partitioned into n separate chiplets, such that the combination of those n chiplets provides the equivalent functionality of the original T-transistor SoC while the sum of the costs of the individual chiplets plus any additional costs for reintegration (e.g., additional packaging expenses) still comes in lower than the cost of a monolithic T-transistor SoC, then a chiplet implementation for this SoC may be worthwhile.

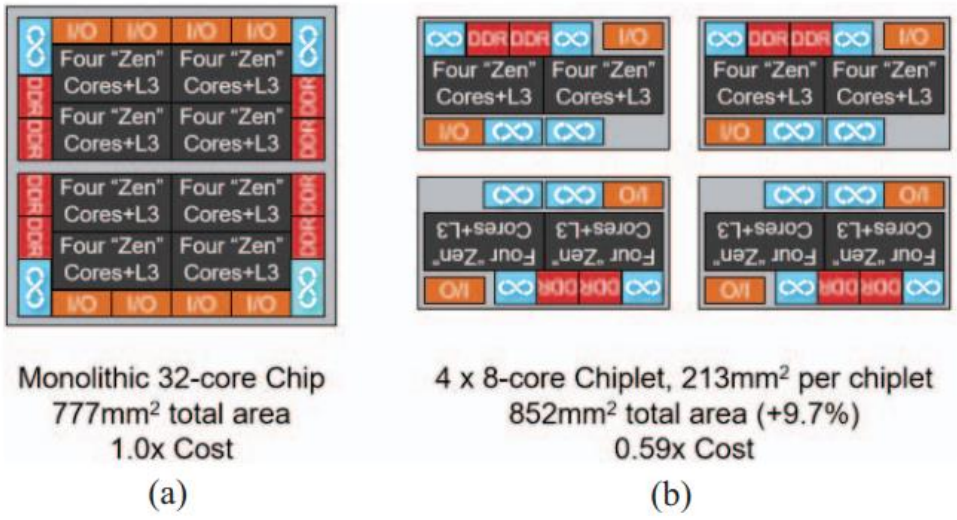


Fig: Hypothetical monolithic 32-core chip compared to an assembly of four eight-core chiplets.

INDEX

EXECUTIVE SUMMARY

TASK 1.1

Covering the technical aspects of the necessity of adapting to Chiplets for JLR as a manufacturer.
We've explained 4 Major applications for the use of Chiplets.

1. ADAS
2. In-Vehicle-Infotainment
3. V2X (Vehicle to Everything Communication)
4. Edge Computing

We have also compiled data from various sources on uses and node sizes that we can divide SoC's into by application. These will ensure that we do not use unnecessarily smaller node sizes for less intensive operations. Therefore, opening up more suppliers and reducing the cost.

TASK 2.1

Packaging technology significantly impacts low-power, high-bandwidth communication. Intel's expandable and rapid system and TSMC's compact low-power solution target different applications. We'll assess why intel's solution outperforms TSMC's for our specific needs. We have also touched upon how JLR can take the lead when it comes to utilizing NPU's such as Mythic AI and Silicon Photonics to get a huge boost in performance.

NVIDIA's NVLink and NVSwitch technologies facilitate swift data exchange and computation collaboration between GPU and CPU processors, significantly enhancing communication speed and data throughput in multi-GPU systems. NVSwitch, an 18-port NVLink switch, enables robust communication between all GPUs in a system. GPUDirect, another NVIDIA technology, minimizes data transfer latency and boosts performance by enabling direct GPU access to data, making these technologies valuable for high-performance computing, AI, and data-intensive applications.

PCIE stands out as the go-to standard for connecting diverse hardware components such as graphics cards, SSD's, and network cards within a computer system. Ucie, on the other hand, specializes in on-chip communication, facilitating high-bandwidth, low-latency connections between various elements on a chip. CCIX is purpose-built to establish cache coherency and high-speed communication between processors and specialized accelerators like GPU's and FPGA's in data centers, while cxl focuses on optimizing performance for data-intensive workloads, such as ai and high-performance computing, by enhancing memory coherence and enabling swift data transfer.

WHY CHIPLETS FOR JLR?

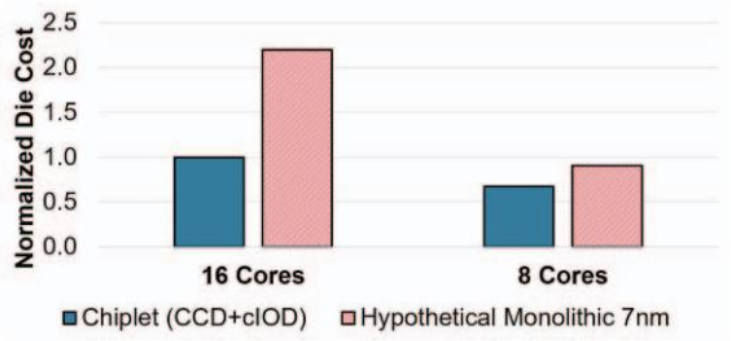
"Different processes can be used for different chiplets, because not all functions need the most advanced (and expensive) semiconductor technology nodes."

"The opportunity to use optimized processes for specific tasks, combined with the sharing of resources such as wiring and cooling infrastructure leads to overall energy savings."

Chiplets are less complex to manufacture than large monolithic chips. That leads to fewer defects, thus higher yields.

CHIPLETS LEAD TO REDUCTION OF COST

AMD presented two slides that detail the extent of cost savings yielded by its bold decision to embrace the MCM (multi-chip module) approach to not just its enterprise and HEDT processors, but also its mainstream desktop ones. By confining only those components that tangibly benefit from cutting-edge silicon fabrication processes, namely the CPU cores, while letting other components sit on relatively inexpensive 12 nm, AMD can maximize its 7 nm foundry allocation, by making it produce small 8-core CCDs (CPU complex dies), which add up to AMD's target core-counts. With



this approach, AMD can cram up to 16 cores onto its AM4 desktop socket using two chiplets, and up to 64 cores using eight chiplets on its SP3r3 and sTRX4 sockets.

Fig: Normalized AMD Ryzen™ processor costs as core counts vary compared to hypothetical monolithic die.

As semiconductors get more advanced, they get smaller. At a sub 10nm scale, foundries must be spotlessly clean. This brings with it manufacturing complexities. Also, the smaller transistors get, the more likely they are to fail.

Chiplets are smaller functional dies that integrate multiple chiplets into a single semiconductor. By giving functions their own circuits (sub-circuits), we can remove design complexity and focus on efficiency.

Take a wafer-scale, massively parallel, SRAM-laden matrix math engine like the one designed by Cerebras Systems,hold it in the air perfectly level

Let it drop on the floor in front of you and then pick up the perfect little rectangles and stitch them all back together into a system.

More precisely, instead of doing wafer scale matrix math units with SRAM, make a whole lot of little ones that have a very low individual cost and a very high yield

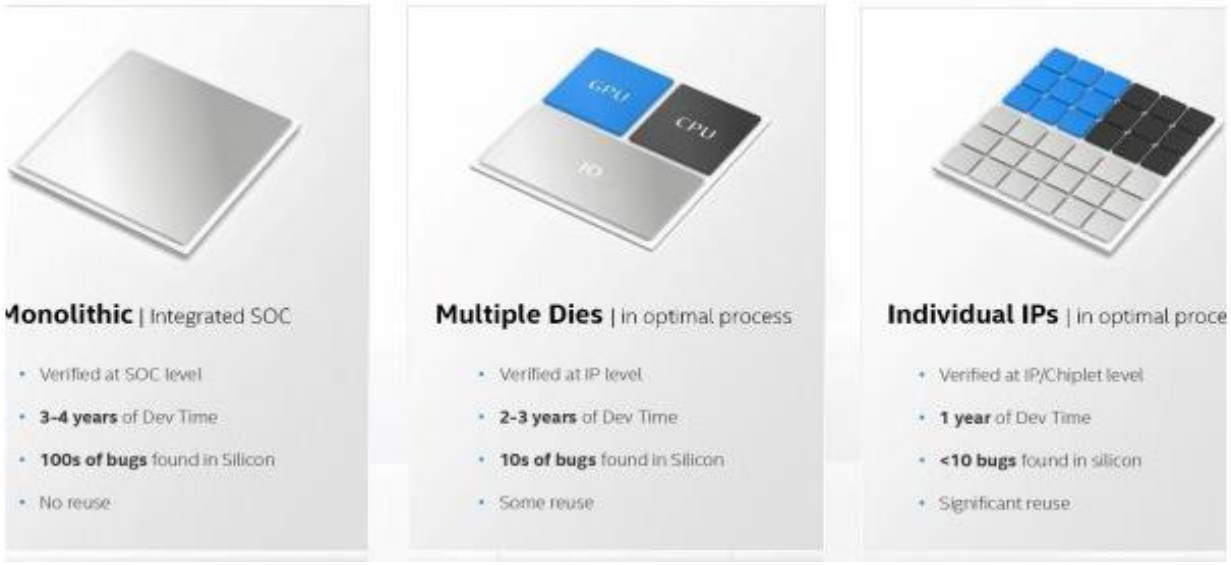
This also pushes that cost down.Then stitch them back together using very fast interconnects.

CHIPLET CLOUD CAN BRING THE COST OF LLMS WAY DOWN:

CHIPLETS SIMPLIFIES THE DESIGN COMPLEXITY AND IMPROVES THE YIELD

We can increase the yield of dies with small transistors by reducing overall size. But as we reduce the size of the die, we have less space for the transistors. Using chiplets maximizes the yield of dies and reduces design complexity, which in turn reduces manufacturing cost.

To give an idea of how much, AMD said chiplet designs can cut costs by more than half. AMD uses chiplet design in its Zen 2 and Ryzen chips. The idea being that taking smaller dies and putting them together improves yield. Intel has a vision of advancing the chiplet design further where instead of multiple dies each block has its own building block.



Note: Chiplet design is already being used by AMD and Nvidia. This means two of the three biggest CPU and GPU companies on the planet are on the chiplet train.

- Standard compute power inside the electronic control unit (ECU) will not be able to process the enormous workloads that come with the ADAS, communication, and entertainment functions of tomorrow's vehicles. Only high-performance compute can rise to that challenge. Such supercomputing can no longer be achieved in one package using monolithic IC design – as the size and complexity would become unmanageable.
- Chiplet design – a modular approach based on heterogeneous integration – allows to scale up the number of transistors and other components without hitting the physical limits of a single chip.

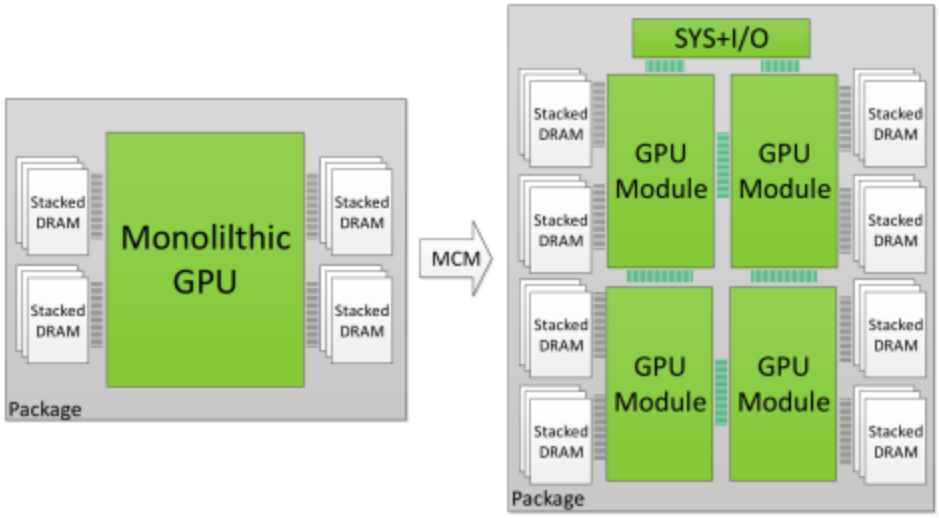


Fig: Aggregating GPU Modules and DRAM on a single package

The

Chiplets are less prone to manufacturing defects due to smaller die size and better wafer utilization at the edges thereby increasing the number of known good die available. With chiplets, the block level yield will improve and thus will allow a higher number of final goods. d I/O bandwidth per GPU.

chiplet-based architecture allows designers to leverage IP without regard to the node or technology on which it is manufactured. Designers can focus solely on their IP or the value-add they bring to the design. These chiplet-based designs can be built on different materials such as silicon, glass,

The Multi-Chip Module GPU (MCM-GPU) architecture is based on aggregating multiple GPU modules (GPMs) within a single package, as opposed to today's GPU architecture based on a single monolithic die. This enables scaling single GPU performance by increasing the number of transistors, DRAM, and I/O bandwidth per GPU.

and even laminate. The result is a high-performance pseudo-SoC built at a lower cost in less time. The reusability of chiplet helps in cost reduction during design and improving yield.

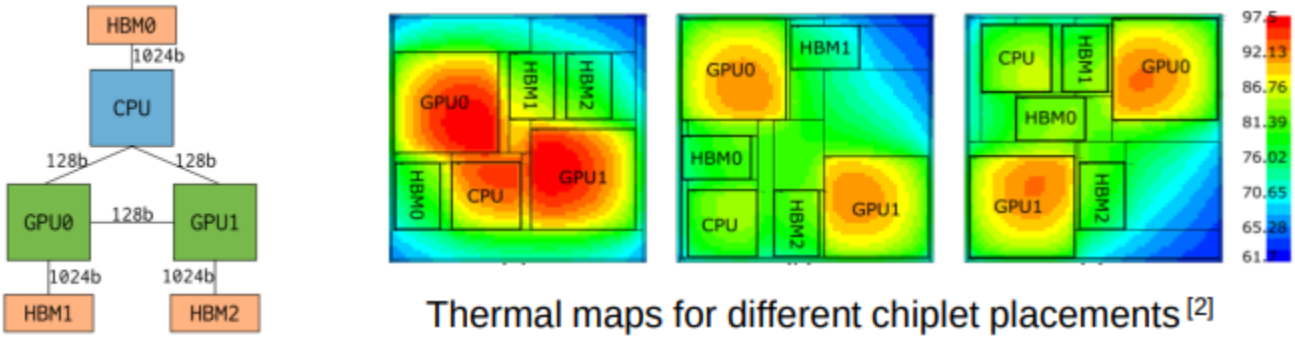
The chiplet-based design distributes heat sources across multiple chiplets. This reduces the localized heat concentration that occurs in monolithic designs and can lead to thermal hotspots. Instead, heat is distributed more evenly across the package, making it easier to manage.

With multiple chiplets, it's possible to design more effective and efficient cooling solutions. Heat sinks and thermal solutions can be strategically placed on the package to dissipate heat more evenly, reducing the risk of overheating and enabling higher performance.

CHIPLETS BETTERS THE THERMAL MANAGEMENT

A standard SOC has all the logic blocks concentrated, which produce the most heat. Chiplet-based design splits them up. If they are spread rather than concentrated thermal management is easier.

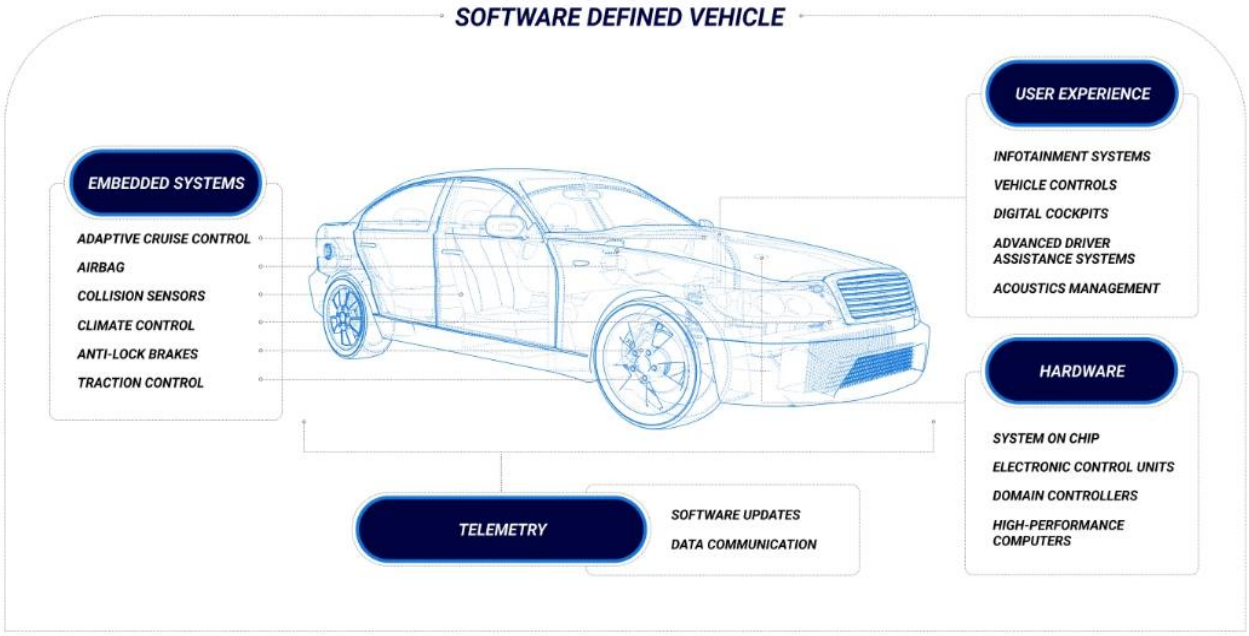
In chiplet-based designs, we can integrate chiplets manufactured using different semiconductor process nodes. For example, we could have some chiplets using a 5nm process and others using a 7nm or 14nm process. This is referred to as Heterogeneous Integration. Lower tech nodes (2nm 5nm etc.) produce more heat than higher nodes. So, with the help of heterogeneous integration of chiplets we can account for higher node chips as well with better thermal management.



Thermal maps for different chiplet placements [2]

SOFTWARE DEFINED VEHICLE

A Software-Defined Vehicle is any vehicle that manages its operations, adds functionality, and enables new features primarily or entirely through software.



Future of SDV:

Software-defined vehicles (SDVs) are orchestrating a paradigm shift within the automotive landscape by metamorphosing cars from electro-mechanical entities into intricate software-centric ecosystems. The future envisions vehicles that seamlessly assimilate, perpetually shaped by the cutting-edge innovations in the digital domain.

Self-driving vehicles (SDVs) signify a synergy between advanced vehicular design and cutting-edge software engineering, catalyzed by key technological advancements which are as follows:

1. Electrification is a key driver in the evolution of self-driving vehicles (SDVs), aligning with the global trend towards eco-friendly solutions. Electric vehicles (EVs) seamlessly integrate with software-driven systems, creating a smooth merger between mechanical design and software functionalities.

2. Connectivity plays a pivotal role in SDVs, with these vehicles becoming integral parts of vast networks. Constant exchange of real-time data enables SDVs to adapt swiftly to changing road and traffic conditions. The hyper-connectivity of today's world enhances the capabilities of SDVs, making them responsive and agile on the roads.

3. Artificial Intelligence (AI) integration sets SDVs apart, enabling continuous learning from driving experiences. SDVs do not just receive manual updates; they learn from every mile driven, contributing to a global data pool. This wealth of data empowers SDVs to make intelligent decisions, anticipate issues, and personalize driving experiences for users.

In this transformative era, collaboration among automakers, software developers, and cybersecurity experts is paramount. Their joint efforts form the core of the automotive revolution, creating a robust web of trust and efficiency. Prioritizing safety and user experience ensures that SDVs become not just technological marvels but also practical and secure modes of transportation, revolutionizing the way we commute.

By 2030, the global automotive software and electronics market is expected to reach $462 billion, representing a 5.5 percent CAGR from 2019 to 2030. In contrast, the overall automotive market for passenger cars and light commercial vehicles (LCVs) is projected to grow at a compound annual rate of 1 percent in the same period—from 89 million units in 2019 to just 102 million units in 2030.

APPLICATIONS

VEHICLE TO EVERYTHING

V2X (vehicle-to-everything) technology enables vehicles to exchange data with their environment, including other vehicles (V2V), pedestrians (V2P), infrastructure (V2I), and networks (V2N). The purpose is to enhance traffic efficiency, safety, reduce pollution, and support advanced driver-assistance systems (ADAS) and autonomous driving.

V2V (Vehicle-to-Vehicle) Communication:

- Exchange of data like speed, position, and direction between vehicles.
- Enables collision detection, coordinated movements, and safe distancing.
- Enhances situational awareness, prevents accidents, improves traffic flow, and optimizes fuel consumption.
- Essential for ADAS and autonomous driving, aiding informed decision-making.

V2I (Vehicle-to-Infrastructure) Communication:

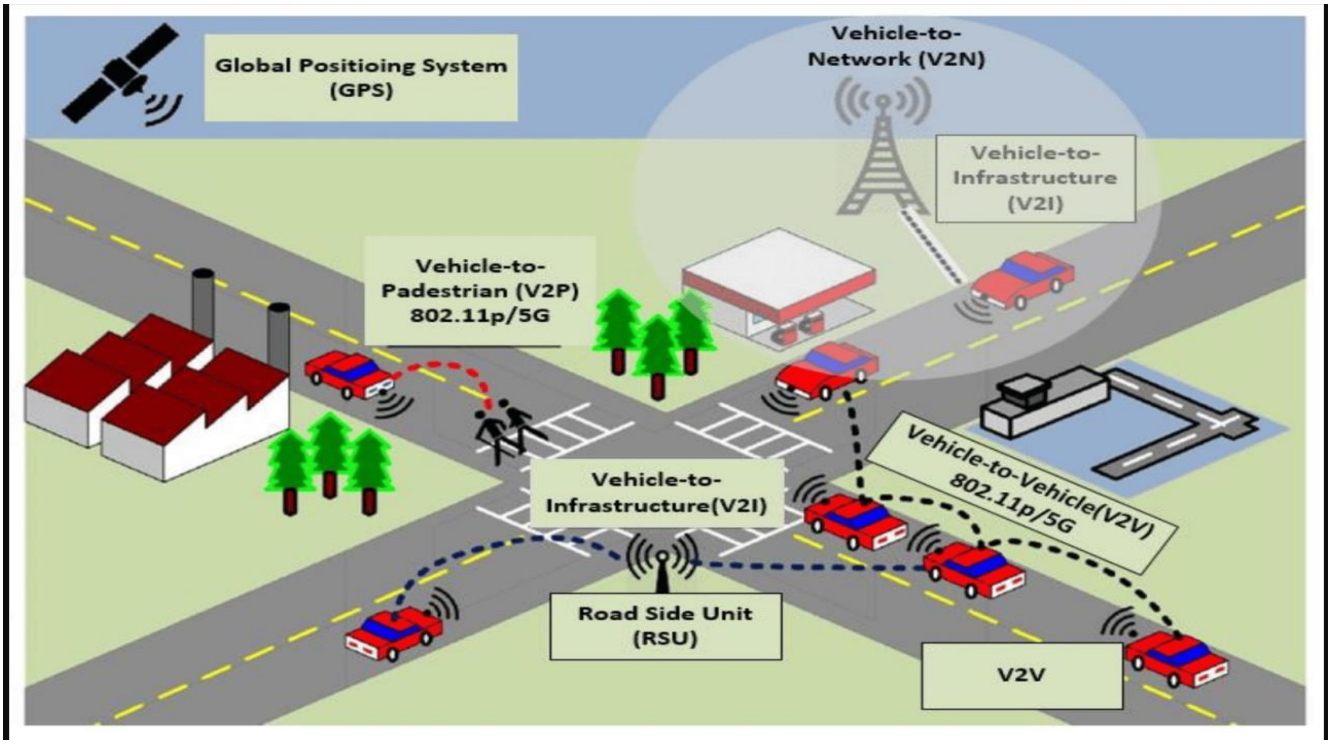
- Interaction with infrastructure elements like traffic signals, road signs, and sensors.
- Provides vital information such as traffic light status, speed limits, road conditions, and hazards.
- Reduces congestion, optimizes traffic signals, and enhances transportation efficiency.
- Contributes to safer navigation and serves as input for ADAS and autonomous vehicles.

V2P (Vehicle-to-Pedestrian) Communication:

- Focuses on interaction between vehicles and pedestrians, cyclists, or vulnerable road users.
- Pedestrian location and movement data transmitted via smartphones or wearables.
- Enables vehicles to identify and avoid potential collisions, ensuring safety for all road users.
- Alerts drivers or autonomous systems about pedestrian movements, allowing appropriate responses.

V2N (Vehicle-to-Network) Communication:

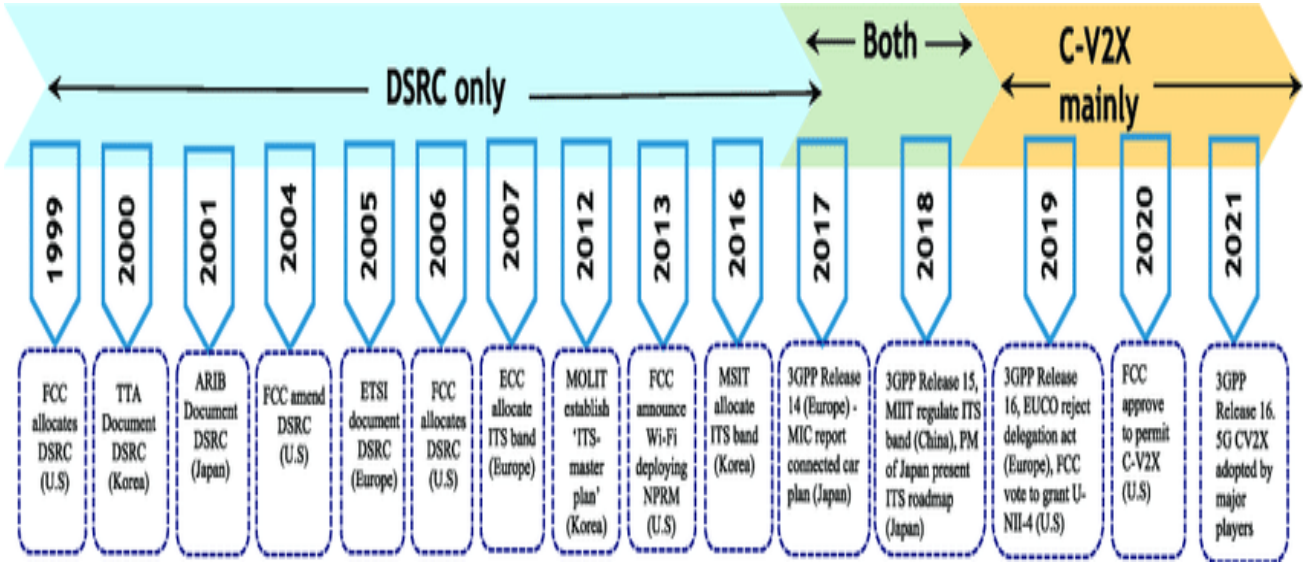
- Connects vehicles to broader communication networks like cellular or Wi-Fi.
- Access to real-time traffic information, weather updates, and route suggestions.
- Facilitates efficient and safe journeys, supports remote diagnostics, and over-the-air updates.
- Integrates vehicle data with public transit systems and city infrastructure, aiding smart cities and connected transportation ecosystems.



V2X'S KEY BENEFITS:

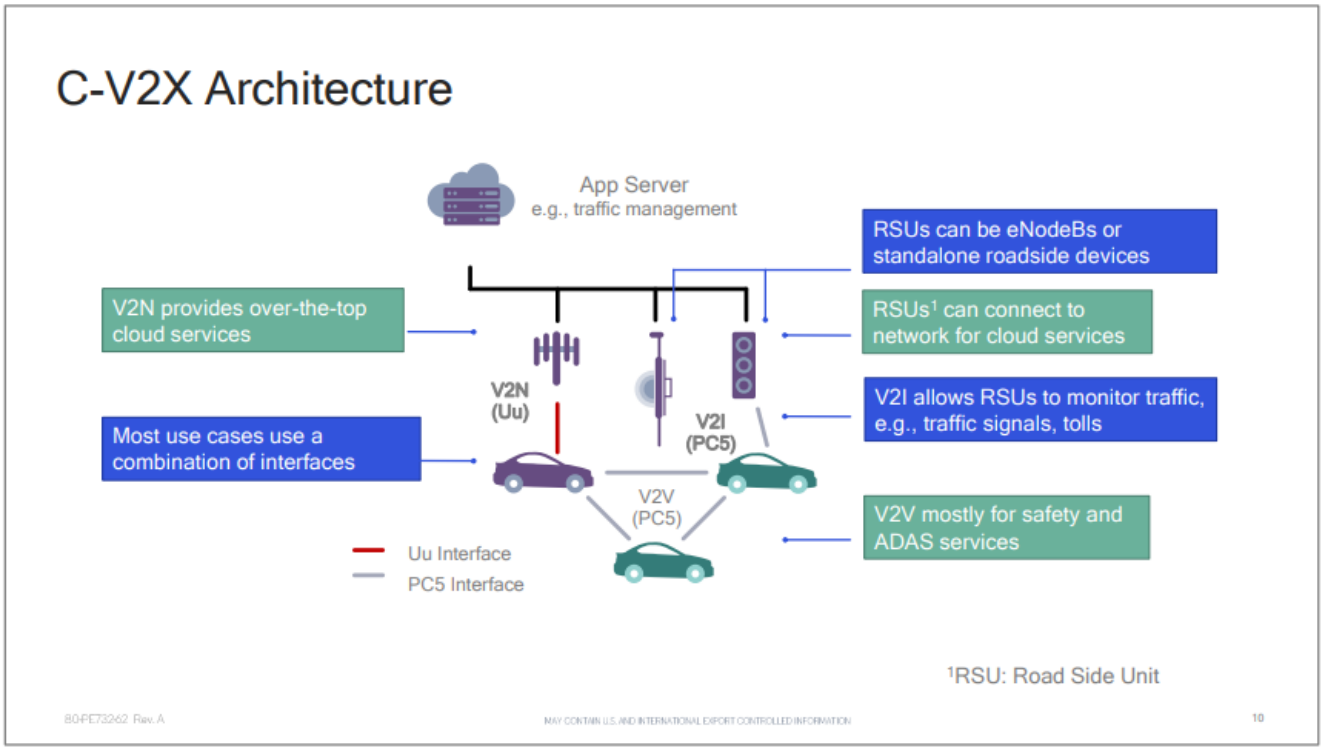
Improved Safety	• V2X enables real-time communication between vehicles and infrastructure, providing information about other vehicles and potential hazards. • Helps drivers avoid accidents, reducing fatalities and injuries on the road. • Addresses visibility challenges, alerting drivers to the presence of vulnerable road users, enhancing safety.
Increased Efficiency	• Reduces traffic congestion and improves traffic flow, leading to more efficient use of road networks. • Lowers fuel consumption and reduces emissions by optimizing traffic patterns.
Enhanced Mobility	• Helps drivers reach their destinations more quickly, safely, and cost-effectively. • Supports efficient route planning and navigation, minimizing travel time.
Improved Transportation System Management	• Facilitates real-time data sharing among vehicles and infrastructure elements. • Enables transportation agencies to manage the road network effectively, respond to incidents promptly, and optimize traffic signal timings.
Better Accessibility	• Enhances accessibility for drivers with disabilities or mobility challenges. • Provides advanced warnings and information to assist drivers in making informed decisions, promoting inclusive mobility.
Reduced Emissions	• Traffic flow management through V2X communication reduces idling and other contributors to automotive-related emissions. • Supports eco-friendly driving behaviors, contributing to a reduction in overall emissions.
Safer Micro-Mobility	• Enables communication with electric scooters, bicycles, and other personal transportation devices, enhancing safety for micro-mobility users. • Alerts drivers to the presence of vulnerable road users, addressing the main cause of micro-mobility accidents: lack of visibility.

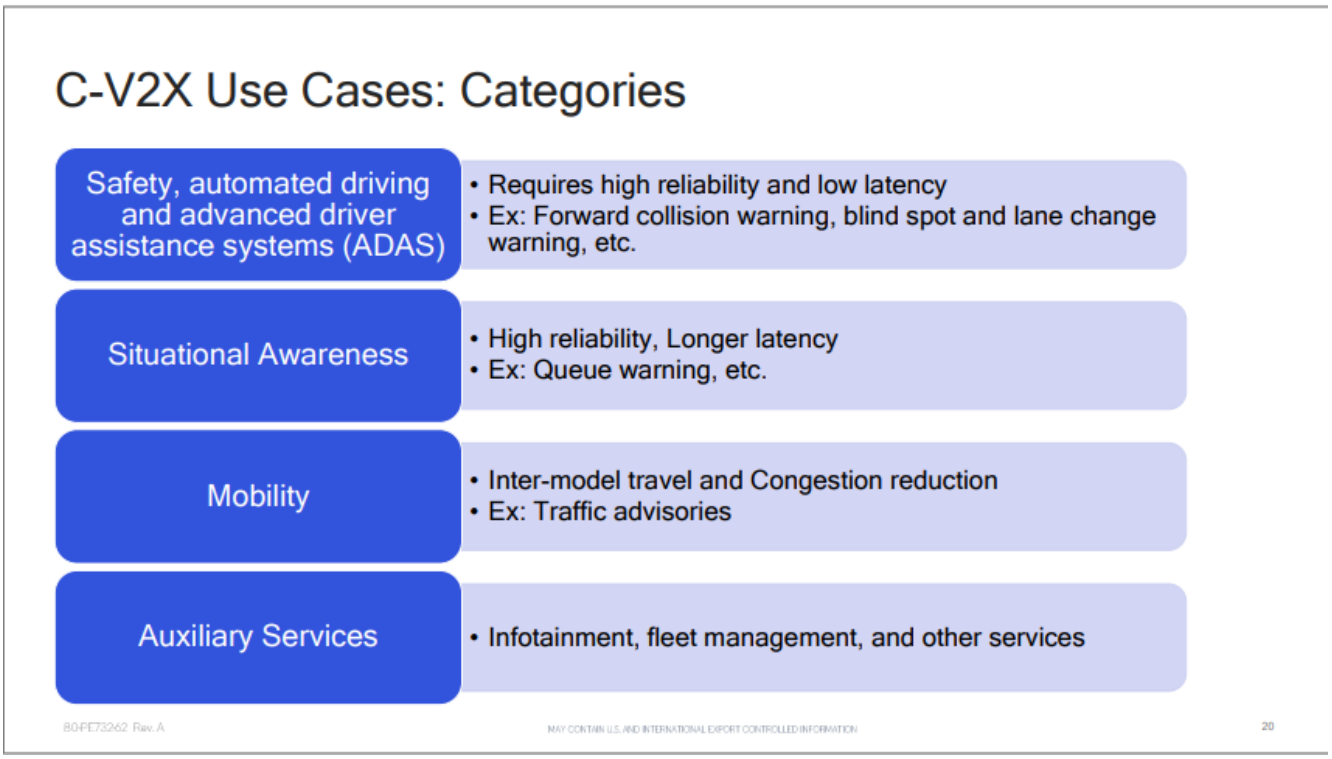
TIMELINE OF V2X EVOLUTION:



C-V2X:

Cellular V2X (C-V2X) is a 3GPP-based global vehicle wireless communication technology that evolves from 4G and 5G. C-V2X includes LTE-V2X and 5G-V2X.





C-V2X SYSTEM ARCHITECTURE AND COMPONENTS:

The C-V2X system architecture employs a layered approach, streamlining communication systems by segregating functionalities into distinct layers:

Application Layer: Hosts various applications like collision avoidance and traffic signal coordination. Utilizes lower layers for data exchange between connected entities.
Transport Layer: Ensures reliable data transmission, managing data packets, flow control, and error detection. Maintains Quality of Service (QoS) for C-V2X applications.
Network Layer: Handles data packet routing, addressing, path selection, and integration with other networks like the Internet.
Data Link Layer: Ensures reliable data frame transmission between nodes, managing Medium Access Control (MAC) protocols and error handling.
Physical Layer: Manages actual transmission and reception of data bits over the wireless medium, including modulation, demodulation, encoding, and synchronization.

KEY COMPONENTS OF A C-V2X SYSTEM:

On-board Units (OBUs):

Hardware devices in vehicles facilitating C-V2X communication.

- Chiplets can enhance OBUs by enabling modular design, customizing components, and ensuring efficient integration, leading to flexibility, scalability, and reduced development time.

Roadside Units (RSUs):

Hardware devices along roads enabling C-V2X communication with infrastructure.

- Chiplets can optimize RSUs by providing specialized components for data processing, ensuring seamless communication, and enabling easy upgrades for evolving infrastructure needs.

Cellular Networks and Core Network Functions:

Provide connectivity between vehicles, cloud-based services, and other connected entities.

1. Chiplets can enhance network functions by offering modular and scalable components for efficient data transmission, resource management, and security, ensuring robust connectivity.

C-V2X Application Servers:

Host various applications and services leveraging C-V2X communication.

- Chiplets can improve application servers by enabling high-performance processing, efficient data analysis, and seamless integration with C-V2X systems, ensuring responsive and reliable services to connected entities.

Chiplet technology enhances C-V2X components by enabling modular, scalable, and customized designs. Integrating chiplets into OBUs, RSUs, network functions, and application servers offers flexibility, efficiency, and adaptability, contributing to the seamless operation and futureproofing of C-V2X communication systems.

ADAS

Autonomous Driving can be split into 5 Levels.

Level 0 – No Automation

1.Used in Conventional Cars

Level 1 – Driving Assistance

1.Present among most manufacturers

Level 2 – Partial Driving Automation

1.Tesla Autopilot
2.Cadillac Super Cruise
3.Hyundai
1.NAVYA
2.Waymo
3.Magna
4.Volvo & Baidu
4.KIA
5.Blue Cruise by Ford (Approved US Highways only)
6.The MG Gloster was one of the first in India to get ADAS Level 2. The MG Astor, the Hyundai Tucson, the Honda City Hybrid and many others also get ADAS Level 2 in India.

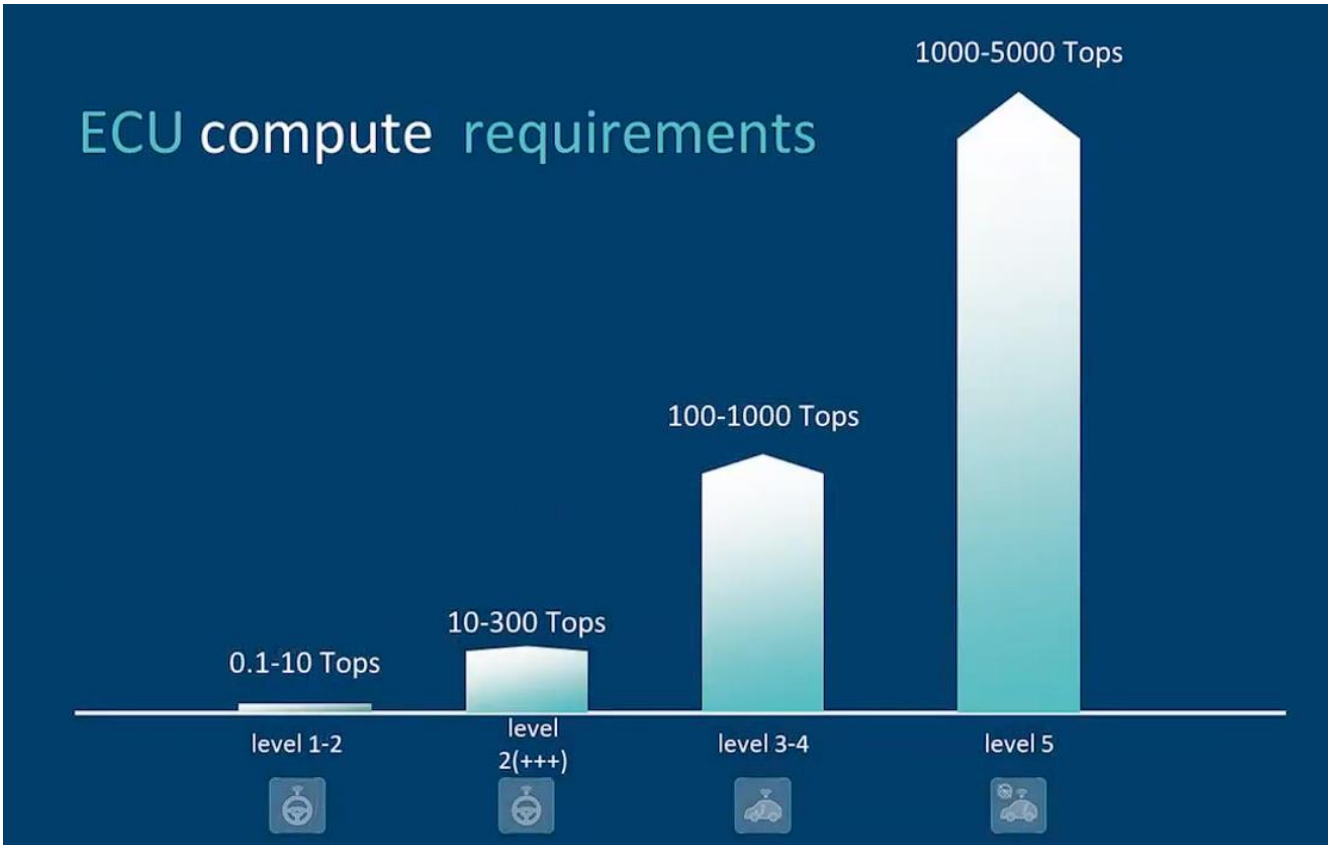
Level 3 – Conditional Driving Automation (Environment Detection)

1.Audio A8L (level 3 in Europe)
2.Mercedes Benz (First to achieve)

Level 4 – High Driving Automation

- Valeo
- Waymo

Level 5 – Full Driving AutomationNot in production, undergoing testing

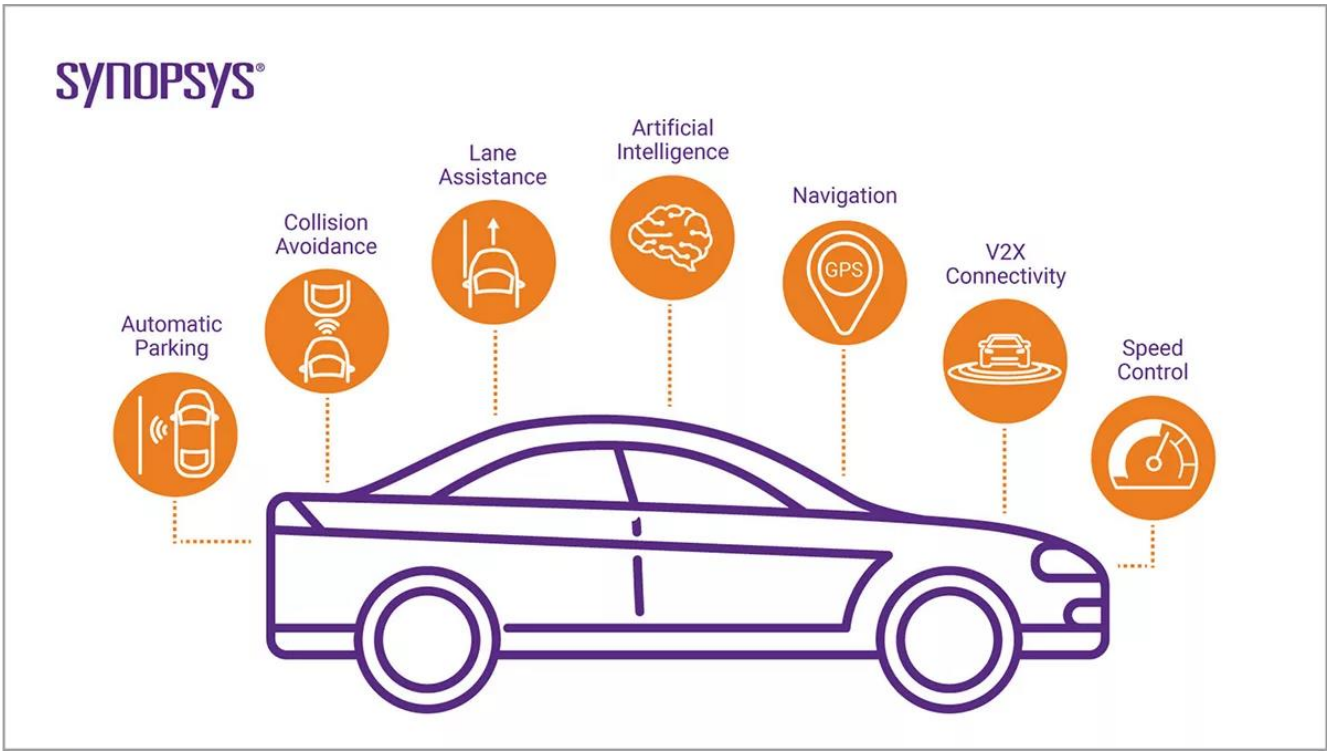


"CHIPLETS ARE FOR THE HIGH-END PREMIUM, WHERE YOU NEED SO MUCH PERFORMANCE AND POWER THAT YOU BREAK OUT OF THE ENVELOPE OF WHAT A SINGLE DIE CAN DO."

- BART PLACKLEY

Going to a higher level of automation brings us to a standstill as we need to make many ethical issues pertaining to prioritizing the driver or the pedestrians, etc.
Also, we must take note that Level 3 is not approved in many markets like the USA.

Upon evaluating the current competition and the feasibility of implementing these levels, we have concluded that it would be appropriate for Jaguar Land Rover to focus on Level 2 (Partial Driving Automation).



LET US DISCUSS THE REQUIREMENTS OF THIS LEVEL IN DETAIL:

Using chiplets can provide scalable architecture from Level 2 ADAS up to Level 5 full autonomy with over 2000TOPS of performance by adding in additional chiplets, but there needs to be software compatibility.

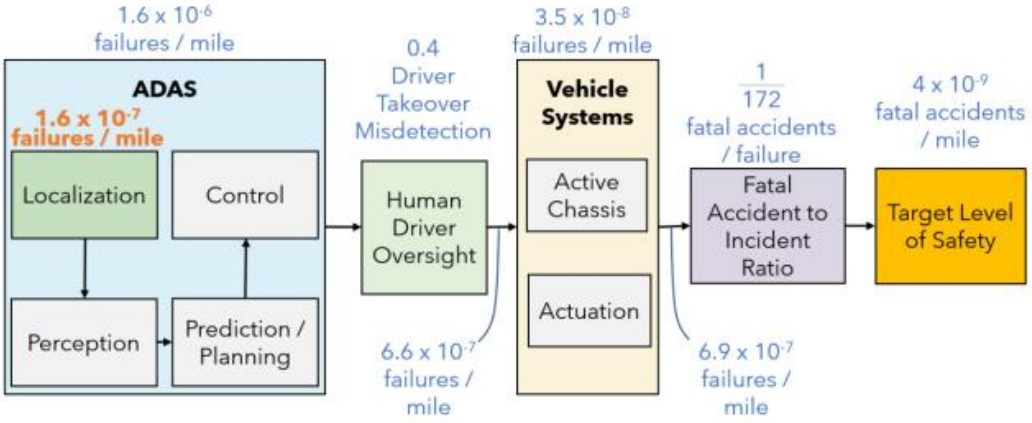


Figure 10: Integrity risk allocation model for ADAS. This includes the human driver in the loop serving an in an oversight role. This shows an example of an ADAS system designed to the level of performance of a human driver in its Operational Design Domain (ODD).

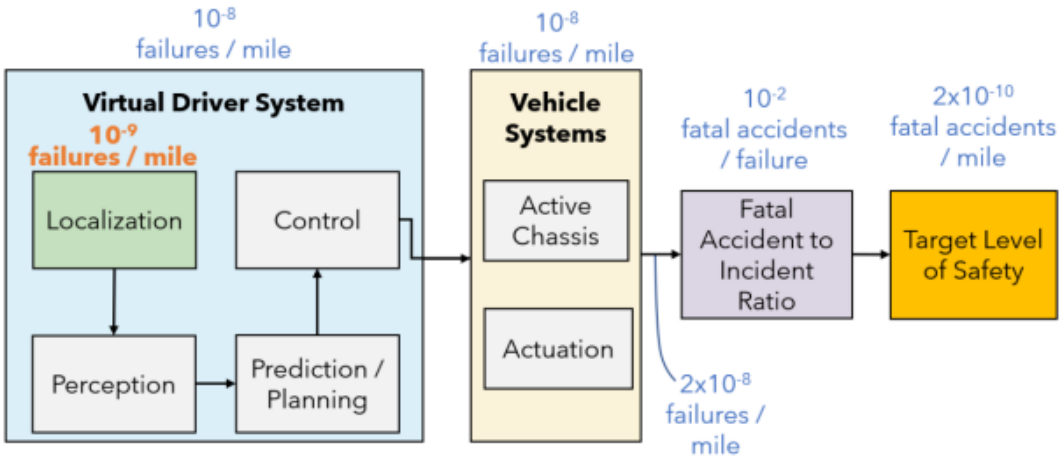


Figure 8: Simplified integrity risk allocation for a Level 4+ self-driving system. This shows the integrity risk allocation of the localization subsystem required to meet the desired Target Level of Safety (TLS) akin to that achieved in civil aviation. (Reid, Houts, et al., 2019).

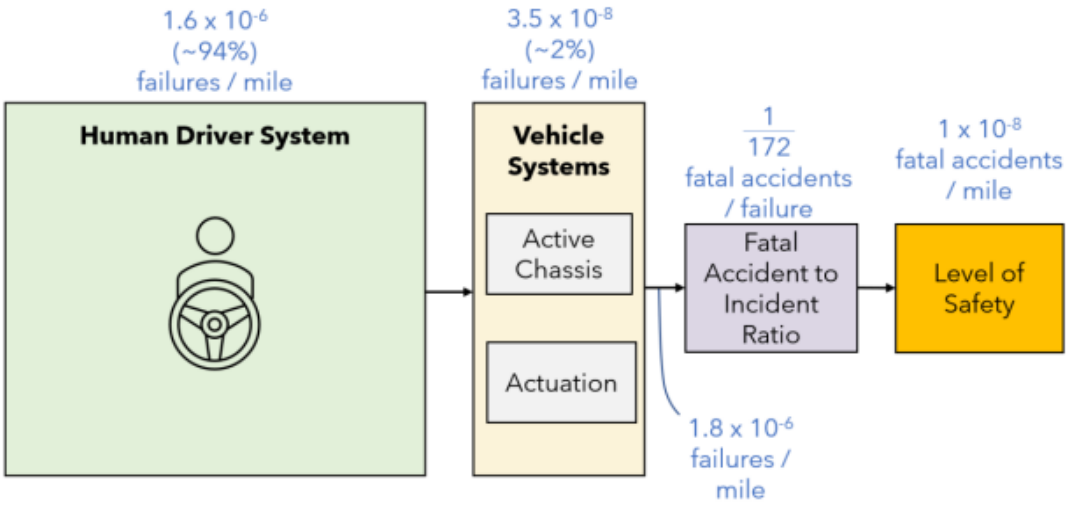


Figure 9: Data-driven model of road vehicle systems today. This includes a model of the performance of both vehicle systems and the human driver in terms of probability of failure per mile.

		Total score	Congestion scene	Special scene	Auxiliary lane change	Curve scene	Human-computer interaction	Automatic parking	Night scene	Rainy scene
1	 Tesla Model 3 v10.2 2020.24.6.4	283	44	59	50	15	twenty one	18	38	38
2	 BMW X5 unknown	208	41	42	34	7	6.5	twenty four	28	25.5
3	 Weilai ES6 v2.6.5	203	39	46	36	14	15	18	35	0
4	 Ideal ONE v1.2.4	196	45	39.5	27	5	17	13	26	23.5
5	 Xiaopeng G3 v2.2.1	171.5	41	33.5	35	3	12.5	25	21.5	0

DIVISION OF CHIPLETS BASED ON FUNCTION:

ADAS COMPONENTS	Purpose	Required Node Size (SoC)	Required Node Size (Chiplet)	Dedicated Node
Camera	Lane Keeping Assistance	14nm	14nm	Hardware Accelerators (GPU, NPU, etc.)
GPS and Communications	Navigation Assistance	14nm	28nm	Communications
Tire Pressure Sensor/Speed Sensor/Ambience Sensors/ Ultrasonic Sensors/ IMU/ Forward Radar	Ensuring safety and Optimal performance, Adaptive Cruise	14nm	40nm	Sensors
Infotainment	Provides Crucial Telemetry information to the user	14nm	14nm	CPU, Hardware Accelerators
Cybersecurity	Ensures Autonomous Driving Capabilities and connectivity is not misused	14nm	14nm	Hardware Accelerators
Power Management	Power Distribution to the Powertrain and the processing components	14nm	130nm	Power Distribution Unit

EDGE COMPUTING

Vehicles equipped with advanced driver assistance systems (ADAS) collect and need a large amount of data to train deep learning and machine learning algorithms that can assist future cars with driving or driving assistance or in autonomous driving. Training such systems typically requires a system capable of capturing and storing real-world data to train the advanced driver assistance system later.

These data are captured and stored by advanced driver assistance systems (ADAS) themselves and are used to develop and improve the performance of future ADAS systems. For example, ADAS uses deep learning or machine learning, and the data collected by an AI edge inference computer is used to train the ML (machine learning) or DL (deep learning) model to increase future reliability. Overall, the more data that is used to train the model, the better the model will perform when exposed to environments and objects it has never seen before. In this regard, the statistics of data becomes an important key matter.

HOW MUCH DATA DO ADAS-EQUIPPED VEHICLES GENERATE?

Autonomous vehicles and vehicles equipped with ADAS can generate anywhere from 1TB to 40 TB of data per day. This number is obvious because vehicles with ADAS are equipped with multiple cameras and sensors that include sonar, ultrasonic, LiDAR, and GPS sensors.

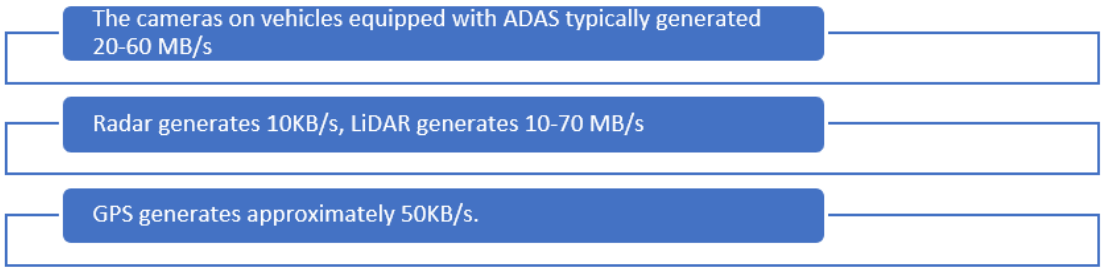


Fig: Data Capture in a typical ADAS System

Upon combining all these numbers, a vehicle is generating 10-150 MB per second of data, which adds up to approximately 8GB of data per minute.

In this regard, the system must store the data onboard the vehicle, because instantaneous submission of the data to the cloud is impossible. It is then obvious that JLR would need an edge

computing solution that has required performance and storage capacity. The need for Edge AI or Edge ML comes into play for the following reason.

- Since the amount of data generated is so huge, it is impractical to keep it in big storage defined inside the vehicle while the issue of high-speed data transfer to storage also remains in place.
- Large amount of real-time data transfer to the cloud is not feasible using the usual cellular communication from the vehicle.
- Not all the data is relevant and good to be sent to the cloud; there could be unusable or bad data that is not suitable or needed for training and ML applications.

It therefore becomes especially important that the data is processed locally and only a useful small amount of data is sent to the cloud for future ML training purposes. Edge AI benefits the system in the given manner-

- Real-time edge processing of the data can reduce and filter the data based on ML, returning only the useful data, which in turn reduces.
  1. Huge storage-dependency and storage device's requirement
  2. time to cloud the data.
  3. cellular bandwidth requirement and connectivity
- Makes the data more secure and close, reducing the chance of data breaches.
- Offers adaptability in data's existence in the vehicle to changing conditions and environments that are not possible when storing all the data in a static manner.
- Direct integration of new features (Edge-trained ML models for autonomy) to the already developed ADAS in the vehicle.

Since the whole process of data retrieving, processing, and filtering requires high power GPU, GPU accelerators, high-speed storage devices and communication, while the data comes from camera, radar, GPS, CAN networks, deploying AI close to the sensors is a good move. The use for chiplets comes in the following way.

- Selective requirement of different node processor chiplets for different tasks optimizes the total resource required.
- The modular nature of the chiplet offers adaptability to the Edge computer and thereby, easy integration of the Edge system to the main central computer.
- Different sensor data and collection-driven selective sensor fusion using chiplets.
- Offers the capability of disintegration after the lifecycle of purpose.

PREDICTIVE MAINTENANCE

An important EDGE application is the predictive maintenance of possible physical faults in the vehicle- brakes, tyres, steering, and underlying machinery that involves the use of real-time sensing and ML classifications to predict when maintenance of a given vehicle part is required. Autonomous vehicles rely on real-time data from sources to assess the condition of critical components, for example, machine learning algorithms can analyze the vibration and temperature data to predict when a component is likely to fail, or using sensor data to check if that subject sensor is working properly or needs maintenance.

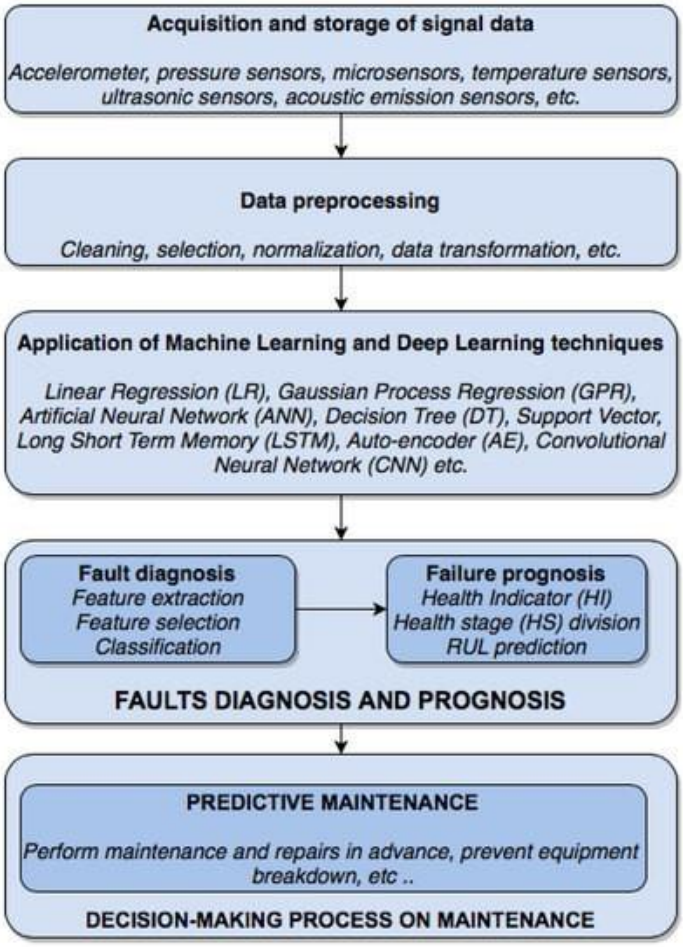


Fig: Predictive Maintenance Workflow

Predictive Maintenance predicts the optimal time point for maintenance actions, with an intent to avoid the premature and costly repair of a system while at the same aiming to ensure a timely repair prior to a failure. Advanced AI methods aim to predict the expected time of a failure, thereby estimating the remaining useful life.

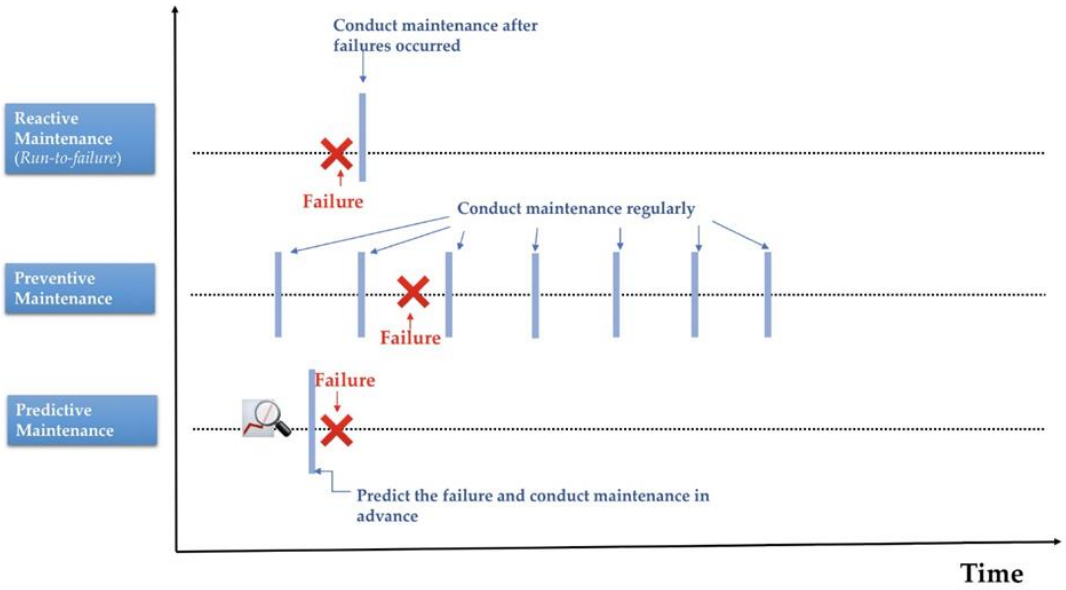


Fig- Comparison- Predictive, Reactive and Preventive Maintenance

The need for Edge-based Predictive Maintenance is the following.

1. 1. The current check for maintenance in vehicles happens on the cloud, leaving a gap period between the test and the fault, leading to
2. a. Increased damage till the time of fault detection than instantaneous maintenance upon capture of failure, way before failure.
3. b. Higher maintenance time and cost.
4. 2. Certain components like Airbags or Braking systems have zero fault tolerance; therefore, predictive maintenance is not a luxury but the only option.

IN-VEHICLE-INFOTAINMENT (IVI)

The infotainment system enables the provision of both entertainment and information for an enhanced in-vehicle experience. The infotainment system covers the main display, multimedia-audio, video, navigation system, software settings, connectivity, and the human-machine interface (HMI), also extends to safety and connectivity within vehicles.

In response to safety concerns, infotainment systems are increasingly incorporating features that minimize driver distraction. This includes improved heads-up displays, gesture control, and enhanced driver monitoring systems and in-cabin sensing to ensure that the driver's attention remains on the road.

Key components/systems inside in-vehicle-infotainment are:

- Head Unit/Central Display
- Instrument Cluster
- In-Cabin Monitoring System
- Telematics and connectivity

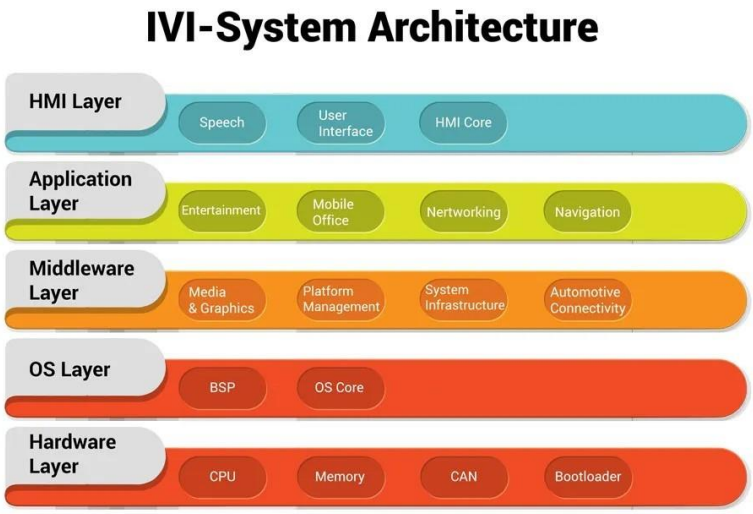


Fig: IVI System Architecture

The Head Unit, or the main human-control interface in a modern car, is a big touch display that is supposed to offer all control to the driver. The display is equipped with navigation, audio-video control, games, vehicle control, speed limit, and more technical and out-of-technical controls. Since the display is a heavy graphic user interface and it is supposed to handle a heavy workload, and the display is intended to keep smooth with good frames per second, operated by a typical OS such as Windows or Linux, it demands a graphics processing unit.

Augmented Reality- There are several car manufacturers that already have integrated augmented reality (AR) into their production cars, including Jaguar Land Rover, BMW, Mazda and Mini, but are in a very initial phase. Using the screen in the center console of the car or the front glass, and essentially a computer within a dashboard, the cars can give the driver extra guidance about their surroundings in real-time, right in front of their eyes. AR is currently capable of displaying the likes of the speedometer, land guidance and directions, and things are about to get a lot better.
Since the Augmented Reality application is a compute expensive thing, it demands a powerful processing unit for real time Matrix Transformation, Pose Estimation, 3D Projection, Coordinate Transformation, Geometric Calculation, Rendering and Graphics.

Fig: Augmented Reality Equipped SDV

In-Cabin Sensing and Driver's activity detection

In the case of an Autonomous Vehicle, the passenger can either be engaged in monitoring the driving or distracted activities such as using handheld devices or reading. Despite self-driving capabilities, it is still within a very reasonable expectation that humans can overwrite or take over the driving task from autonomous vehicles in exceptional situations or emergencies. Since driver decisions and behaviours are essential factors, it is worthy installing driver activity recognition systems for the additional margin of safety.

Human-Machine Interaction for Autonomous Vehicles: A Review | SpringerLink

Fig: In-Cabin Sensing

Fig: In-Cabin Sensing[2]

Since there are multiple applications that demand different hardware resources, chiplet offers the functionality to split one hardware to the dedicated systems, and can also allow for a micro-zonal architecture inside IVI, for there is huge diversity in the IVI applications, which remains not possible without the Chiplet game.

Current Market

Market Size- The automotive infotainment market size was 27.2 billion USD in 2022 and is expected to be 48.2 billion USD by 2030.

In 2025, the global automotive AR and VR market is forecast to reach about $673 billion, up from $0.21bn in 2017. That's a 320,000% increase in investment in only 8 years. Reference

Major Players- An advanced infotainment system is powered by a strong processor. Some of the top in-car platforms on the market include:

4. NXP-MX255 applications processor
- Samsung Exynos Auto V9
- Qualcomm® Snapdragon™ 600 processor (Eragon 600 SoM)

For example, the Samsung Exynos Auto V9 powers the Audi's infotainment system, and that is developed on an 8 nm processor.
The concept of chiplets into IVI is already there.
Mediatek and Nvidia collaboration- Through this collaboration, MediaTek will develop automotive SoCs integrating a NVIDIA GPU chiplet with NVIDIA AI and graphics IP for smart cabin solutions and infotainment, will run NVIDIA DRIVE OS, DRIVE IX, CUDA and TensorRT software technologies — delivering a full range of AI cabin and cockpit functions with quality graphics, AI, safety, and security features.
MediaTek partners with NVIDIA to provide product roadmap for connected cars, ET Auto (indiatimes.com)

Jaguar Land Rover introducing chiplet into infotainments add values-

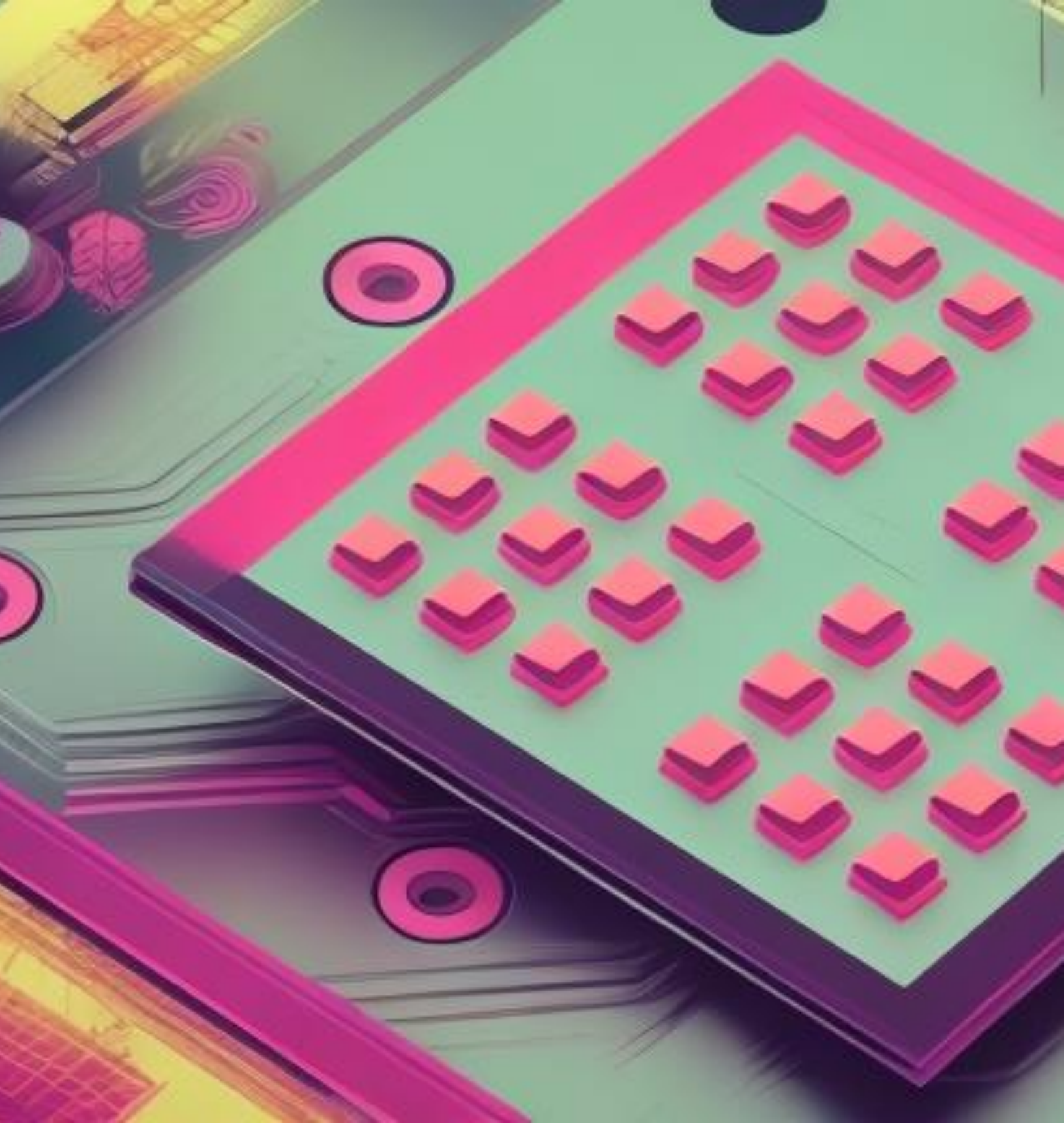
→ 1. Since chiplets technology is unsaturated for infotainments, early introduction of chiplets would give JLR competitive gains.
→

→ 2. Land Rover has a central display sizing more than 25 cm and is supposed to render high quality motions and graphics, which needs a low sized node processor(8-10nm). Disintegrating traditional Chip optimizes the utilized resource.
→

→ 3. Chiplet based architecture allows for micro zonal architecture inside the IVI, causing a lesser complex system with dedicated chiplet based on requirement for each zone.

4. A common Edge based computing solution is workable for both Machine Learning application and general computation.

→ 5. Augmented Reality is a new and exceptionally compute powered thing and requires more resources such as a hardware accelerator beside the normal processor. An SOC does not allow experimentation and needs a single deterministic design, far from the possibility of variables, while chiplets offer options.



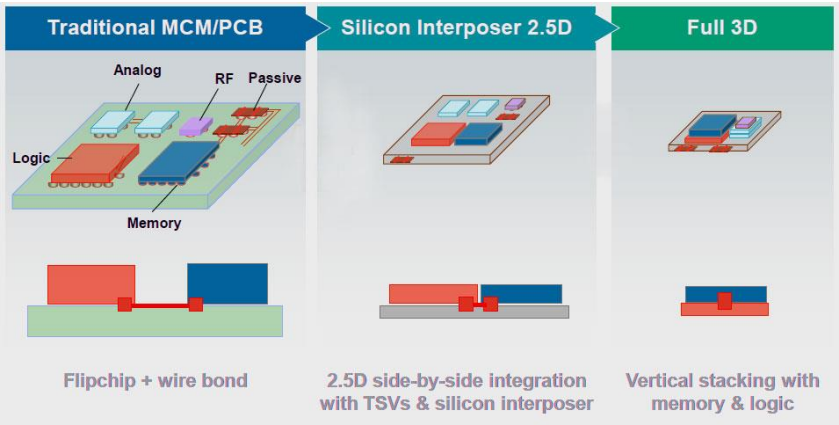
INTERCONNECTS

BACKGROUND OF ADVANCED INTERCONNECT TECHNOLOGY

THE FUTURE OF INTEGRATED ELECTRONICS IS THE FUTURE OF ELECTRONICS ITSELF. THE ADVANTAGES OF INTEGRATION WILL BRING ABOUT A PROLIFERATION OF ELECTRONICS, PUSHING THIS SCIENCE INTO MANY NEW AREAS.

- GORDON MOORE

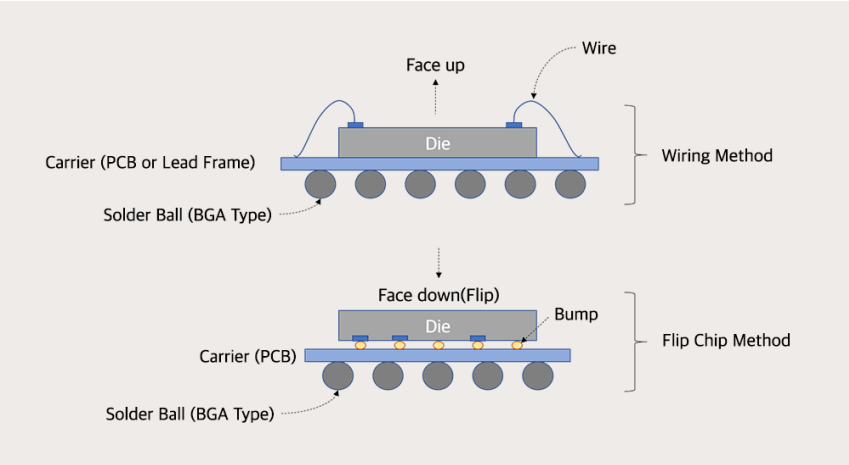
With the increasing need for high performance computational units coupled with the growing hurdles discussed in previous sections the Multi Chip Module is seen as the path forward. The communication between the chips are seen as the major bottlenecks in this approach and prior to discussing the available technologies we need to cover the basic terms and the development path the industry has taken. Packaging technology is one of major players in determining the performance, reliability, and scalability of the chiplet based SoC design.



WIRE BONDS AND FLIP CHIP

The traditional MCM technology developed for years, wire bonds and flip chips have been used to connect between the chip and circuit board. The key distinction being that, the interconnection is not between dies but die to package or substrate. While the most developed and cheapest, they have limitations in high performance and high power applications.

Based on the internal structure, we classify it as wire bonds and flip chip. In wire bonds, the dies are placed face up, allowing wiring to happen on the upper surface. As indicated in the figure, this makes the wires longer compared to flip chip wires; increasing signal travel length.



In flip chip on the contrary, the dies are placed face down on the top of metallic bumps to allow for wiring. This design allows for shorter wire lengths, reducing signal travel time at the expense of manufacturing difficulty and cost. This elementary technology has numerous disadvantages compared to 2.5D and 3D interconnect technology:

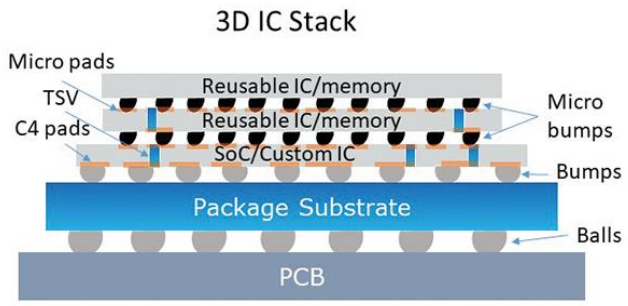
- Higher resistance compared to direct interconnect technology can limit use in high performance applications.
- Flip chips are costly for the stress and reliability concerns that accompany flipping the chip and creating solder connections.
- Limited integration as they are typical for chip to package connections. 2.5D and 3D interconnect technology offer better use of heterogeneous chips.

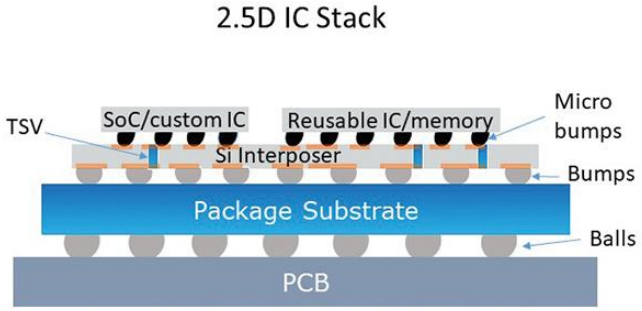
3D INTERCONNECT TECHNOLOGY

3D ICs can be divided into 3D Stacked ICs (3D-SICs), which refers to stacking IC chips and interconnecting them with TSVs; and true 3D ICs, which use fab processes to stack multiple device layers on a single chip, which may or may not use very-fine-pitch TSVs to form the interconnect.

Through silicon vias (TSVs) are holes created in a silicon wafer using an etch process. Interconnects are formed by filling TSVs with a conductive material, such as copper, tungsten, or polysilicon (figure 2). The main advantage of TSV interconnects is the shortened path for the signal to travel from one chip to the next, or one layer of circuitry to the next. This allows for reduced power, and the ability to increase interconnect density, thereby increasing

functionality and performance. TSVs are not 3D ICs all by themselves. Rather, they are the building blocks that enable 3D ICs.





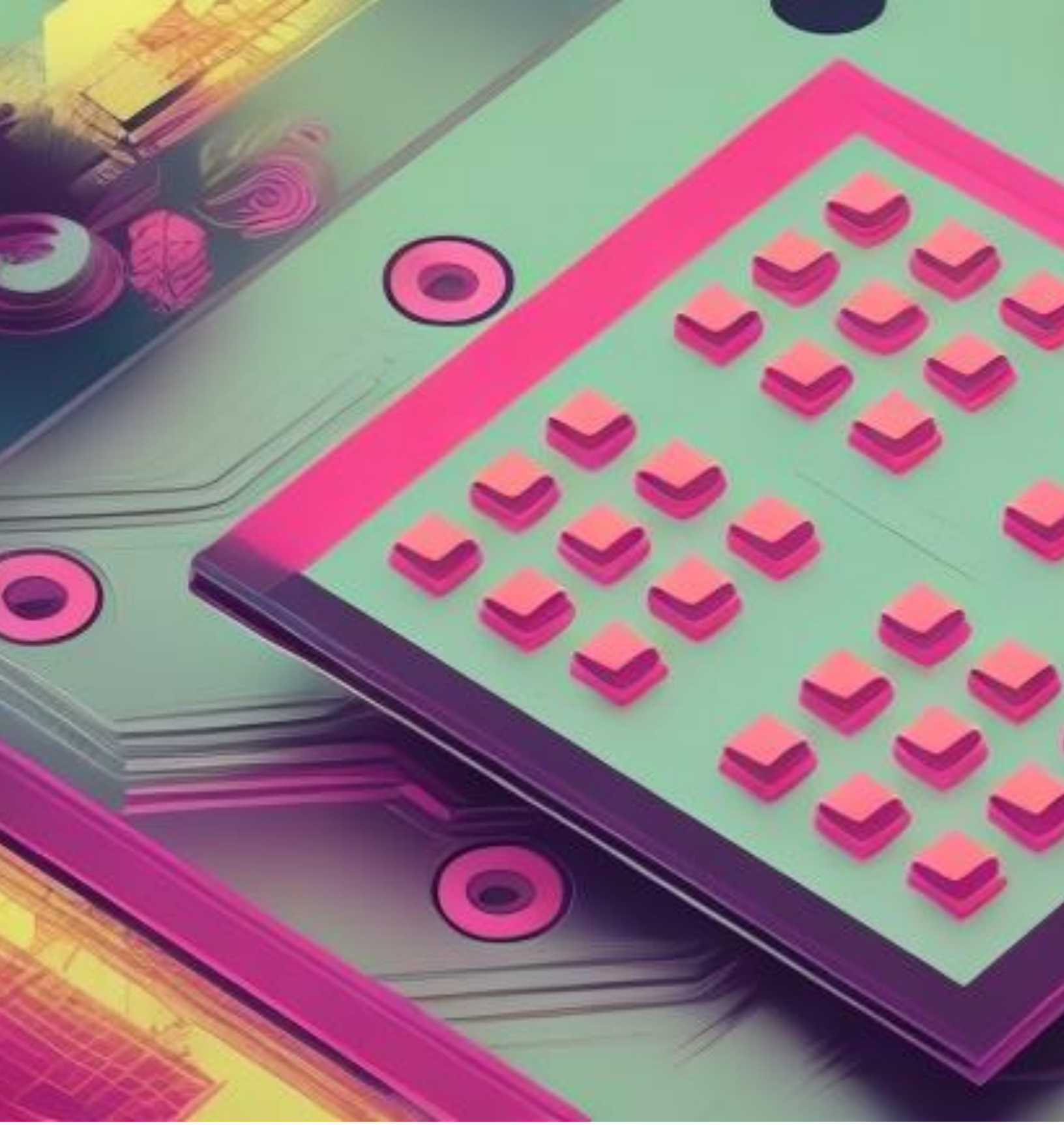
2.5D INTERCONNECT TECHNOLOGY

The naming of the technology gives great insight into the characteristics of the concept. Having experienced the advantages 3D packaging offered, Chip designers discovered that by arranging bare dies next to each other on an interposer instead of stacking them vertically, they could achieve many of the advantages of 3D integration. When the spacing is very fine, and the connections are short, this configuration can be packaged as a single component, offering improved size, weight, and power characteristics compared to a similar 2D circuit board assembly. This midway point between 2D and 3D integration playfully earned the moniker "2.5D," which has since become widely adopted. There are additional advantages that 2.5D packaging offers keeping our case use in mind:

Splitting heat producing dies and spacing them adjacent to each other, rather than being stacked, allowing easier thermal management.

Heterogeneous dies are accommodated easily compared to 3D packaging

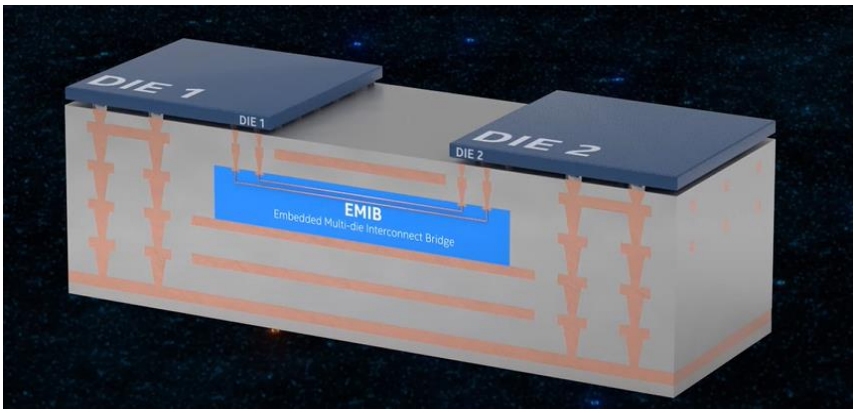
Upgrading the 2.5D assembly is easier. Allowing accommodation for chiplets from numerous suppliers and room for change, as the chiplets require no modifications (in 3D IC Stacking the dies are modified with space for TSVs).



PACKAGING

EMIB

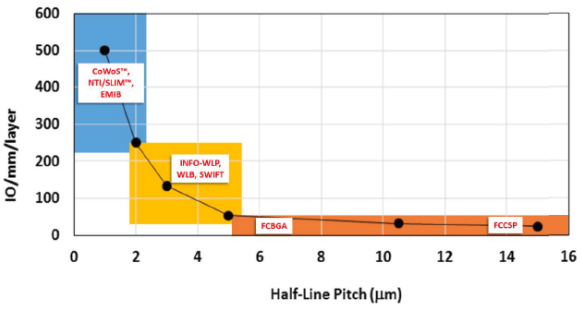
The Embedded Multi Die Interconnect Bridge (EMIB) was first proposed in the mid-2000s, EMIB is unique in that it is the only packaging technology that offers localized high-density wiring, whereas the bulk of the package interconnects are still the same traditional organic package interconnects. The basic concept is that it uses thin pieces of silicon with multi-layer BEOL interconnects, embedded in organic substrates, to enable localized dense interconnects. A very thin silicon bridge is embedded within the organic packaging and connected to flipped chips.

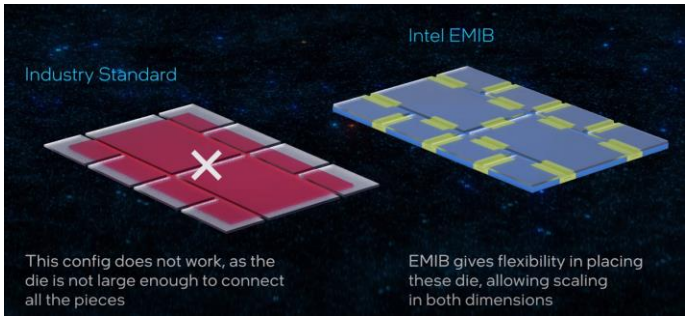


KEY DIFFERENCES

Linear interconnect escape density or IO density (IO/mm/layer) is a key metric used to compare capability envelopes of different packaging technologies. IO/mm/layer is the number of wires escaping per millimeter of die edge for each layer of the package. The only comparable technology offered is CoWoS, SLIM and EMIB.

The key





distinction between these technologies and EMIB is that the former are interposer baked technology. All dies are to be placed on the interposer layer, and the entire chiplet based system must rest on the area of the interposer, limiting the size of the chip. EMIB offers more expandability where there is no limitation in the overall size. Another implication is that since the overall size is not limited (as it is with traditional interposer based 2.5D designs) there are no

limitations on the combination of chiplets to be placed in the limited interposer dimensions. Allowing freedom of integrating numerous combinations of chiplets.

The downsides of this technology, which seems to be ideal up until now, is regarding the difficulty in embedding the bridge with significant yield. Since multiple bridges must be added to the substrate, the probability of all bridges to be manufactured without error brings overall yield down. Even if the yield to embed one bridge is 99%, the yield for embedding 16 bridges would become 85%.

Intel has already developed products with EMIB, most notably the Stratix FPGA and the Ponte Vecchio supercomputer.

Interconnect density: 300 IO/mm/layer

Maximum reach of 15mm at 1Gb/s and 5mm at 5Gb/s

A data rate of 25 Gb/s with 40 I/O/mm

Two routing layers can achieve a 1-Tb/s/mm BW density.

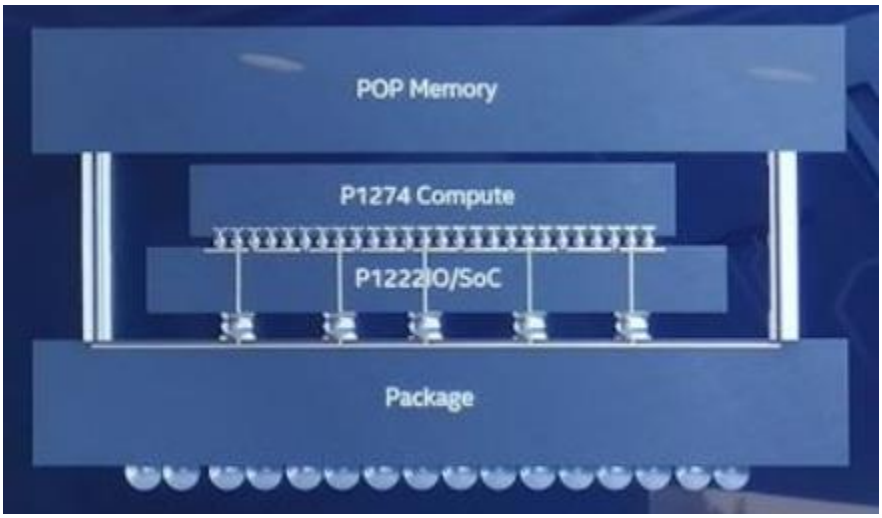
I/O Bandwidth: 1 TB/s using 4096 data signals at the rate of 2 Gb/s and 14mm die edge'

The first generation had a bump pitch of 55 µm, 2nd and 3rd gen offered a 45 µm and 40 µm pitch respectively

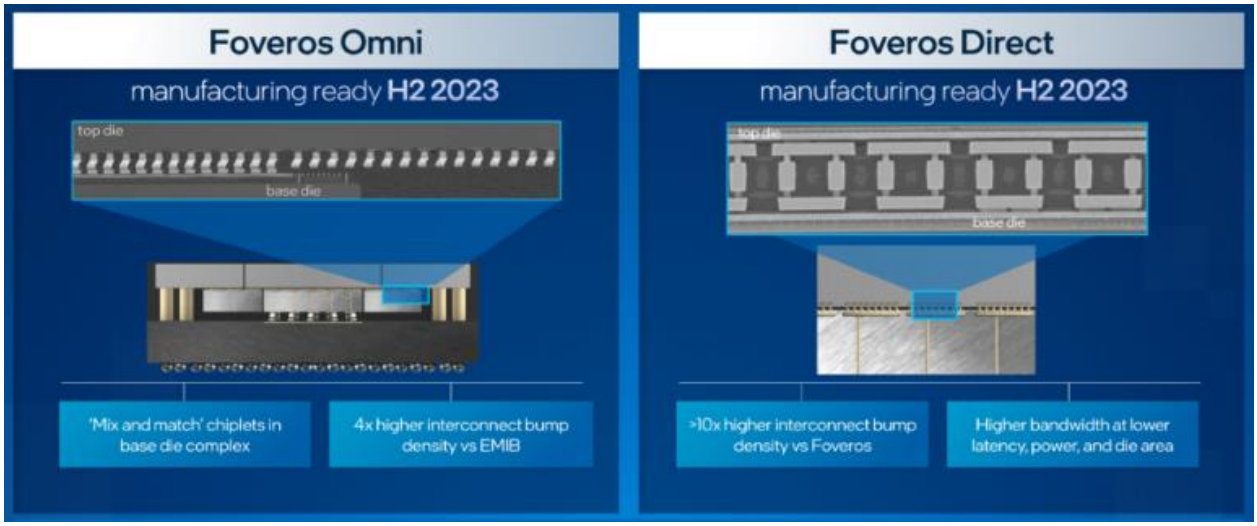
FOVEROS

Intel introduced its die-to-die stacking technology named Foveros in 2019 with Lakefield. Foveros is very similar to interposer based interconnect technology with one major difference that makes it stand apart, the base die which acts as an interposer layer is now an active circuit.

Keeping the Lakefield design in mind, the top die was made with Intel's 10nm process handled the graphics and contained the cores, the base die containing PCIe lanes, USB ports security and everything low powered related to IO was built on 22FFL low power node, a 22nm



based technology. By now placing one die on top of another we entered 3D stacking, it brought shorter data paths, better latency, and lower power loss. Thermals was an issue which was addressed by having the more heat dissipating die on top and a low power die with minimum logic at the bottom. A key point to note in the design is that the upper die must be smaller than the lower die. Having said that, the issue is enabling the top compute die to have power for its logic – this involves large power through-silicon-vias (TSVs) from the package up through the base die into the top die, and those TSVs carrying power become an issue for localized data signaling due to interference caused by high currents.



FOVEROS OMNI

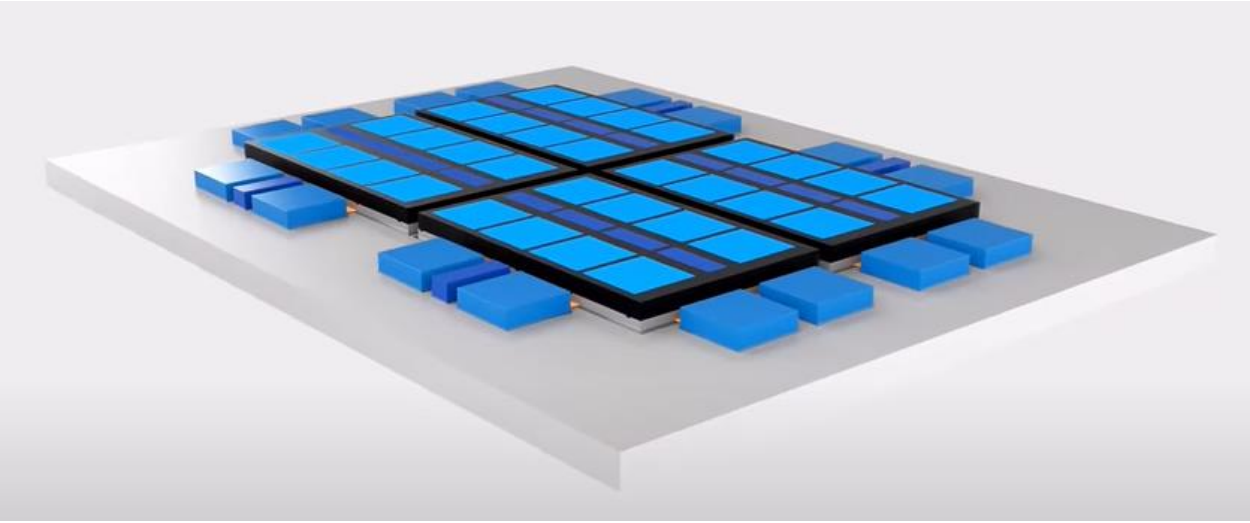
The major advancement in the later generation is that the top die can now be bigger than the lower die. Its goal was to address the power supply issue discussed above. The issue is addressed by allowing the cantilevered section of the larger top die hence keeping the power carrying TSVs outside of the base die. By moving the power TSVs outside the base die, this also allows for a die-to-die bump pitch improvement. Intel is citing 25 microns for Omni.

FOVEROS DIRECT

As these technologies are growing copper and depositing tin solder, it gets difficult to scale them down, plus there is also the power loss of the electronics transferring into the different metals. Foveros Direct gets around this problem, by doing direct copper-to-copper bonding. If one piece of silicon is lined up directly with another, then there is little-to-no need for extra steps to grow copper pillars and such. The issue comes with making sure that all the connections are made, ensuring that both the top die and bottom die are so incredibly flat that nothing can get in the way. Also, the two pieces of silicon must become one, and are permanently bonded together without any way of coming apart. Foveros Direct is a technology that helps Intel drive the bump pitch of its die-to-die connections down to 10 micron, a 6x increase in density over Foveros Omni.

CO - EMIB

The future of Intel's products lies in the concept of combining the two interconnection processes to have a best of both worlds' solution to the packaging. The organic package is developed with EMIB embedded, onto which full Foveros stacks and auxiliary chiplets like high bandwidth memory are brought together.



COWOS

The pursuit of high bandwidth and low signal latency interconnects become increasingly critical in high density heterogeneous integration. Chip on wafer on substrate is a TSV interposer based multiple die wafer-level system integration (WLSI) to achieve higher system performance The first-generation CoWoS products are high-end field programmable gate array (FPGA) with a Si interposer die area up to ~800 mm2, very close to the maximum area of a reticle field. The limitation for interposer die area was soon extended and a CoWoS-based FPGA product with interposer size ~1200 mm2 has been proposed. In 2016, with the advent of second-generation high bandwidth memory (HBM2), CoWoS capability was extended to include logic and third-party memory dies. TSMC characterizes the following advantages of the interposer based system over stacking chips:

1. The mature Si technology used in the interposer does not take extra effort for TSV characterization after a new technology node is available. Improving time to market
2. The top dies are independent function units of Knowns Good Die (KGD). Stacking requires modification, often thinning of the die.
3. Vertical stack sees challenges in heterogeneous die integration as they are often of different sizes.
4. Existing thermal management systems currently used in flip chip technology can be utilized whereas stacking will require new thermal solutions to be adapted.

Increasing the interposer die area is an obvious option in addition to vertical stacking. To extend the interposer die area limitation beyond a full reticle field, we developed the second-generation CoWoS (CoWoS-2) technology. A two-mask stitching photolithography was developed to fabricate the ultralarge interposer. 1200 mm2 interposer area, which is about 1.5 times of a full reticle size, is set as a reasonable goal of CoWoS-2.

While continuous increase of Si interposer size is still an option for next generation CoWoS scaling up to 4x (~3300 mm2),challenges emerge from productivity and reliability aspects. a monolithic Si interposer at such a large size brings up yield concern, especially the gross die count per wafer is dramatically decreasing beyond 3x. Hence, CoWoS-S scaling towards fourtime reticle size (~3320 mm2) or beyond is extremely challenging in terms of production and reliability.

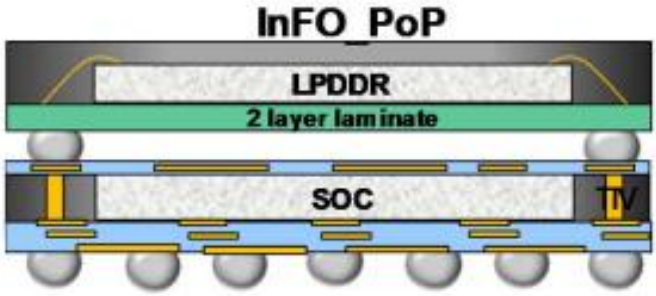
Given the target HBM2 memory bandwidth of 256 GB/s, a CoWoS-2 SiP can deliver up to 1 and 1.5 TB/s total memory bandwidth with four and six HBM2 on an interposer, respectively.In a

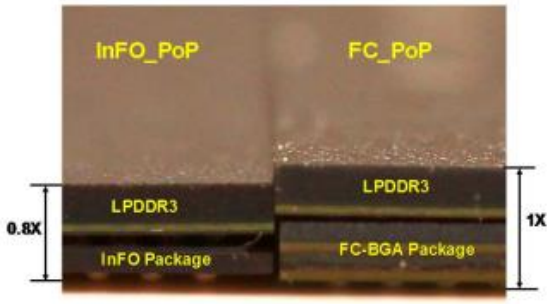
typical CoWoS process, top dies of known-good logic SoC and HBM are integrated in a side-by-side manner on a Si interposer wafer through μ-bumps at a pitch around 30 to 60 um.

INFO

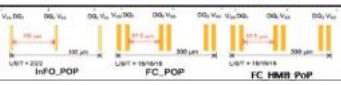
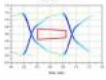
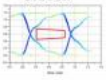
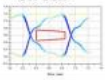
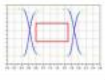
Integrated fan-out (InFO) wafer level system integration (WLSI) technology has been developed to integrate application processor chip with memory packages for smart mobile devices. With its high flexibility and strong capability of multi-chips integration for both homogeneous and heterogeneous sub-systems, InFO technology provides a system scaling solution. InFO is developed keeping smartphones and other hand held devices in mind, prioritizing the overall thickness of the combination. Figure shows the size advantage InFO provides. Another key point to remember is that this is made for memory and processor interconnection, not general purpose. TSMC develops the method to be compatible with numerous memory suppliers.

The performance of InFO is quantified in comparison to related technologies, Flip chip PoP (Package on Package), FC_HMB_PoP and 3D IC using TSV.





As quantified above the InFO_PoP implementation has better signal integrity compared to flip chip implementations and much better thermal performance compared to its immediate competition. This along with its better sizing and multi supplier compatibility emphasis its use in the smartphone industry.

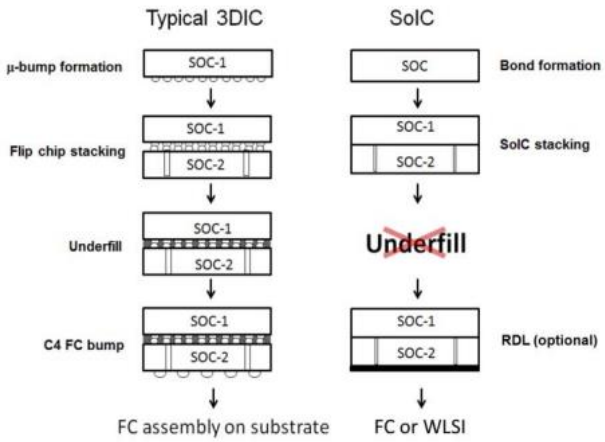
Package Type	InFO_PoP	FC_PoP	FC_HMB_PoP	3DIC
Test vehicle				–
Data IO # per Memory PKG	64	64	64	512
Signal Integrity (Eye height)	1x 	0.833x 	0.758x 	– 
Memory I/O Power Consumption	1x	1.002x	1.005x	0.822x

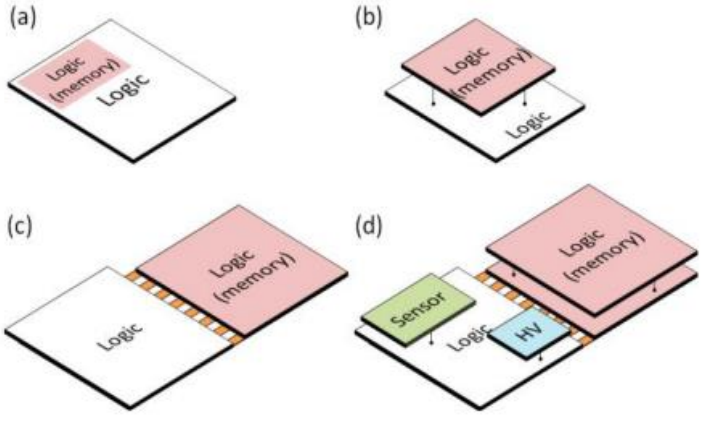
Steady-State Package Thermal Performance		Memory-Stacked Package Schematic			
Items	Leg	InFO_PoP	FC_PoP	FC_HMB_PoP	3DIC
Junction Temperature (SoC/DRAM) T_j, max (°C)	–	84.1/80.9	95.6/91.5	93.3/91.9	97.8/97.8
Thermal Resistance θ_{JA} (°C/W)	–	12.5	15	14.5	15.5
Norm. Leakage Current at T_j, max	w/o Board-Level UF	66%	100%	93%	109%
	w/i Board-Level UF	53%	76%	71%	82%
Max. Allowable SoC Power (W)	No Lid, w/o BL UF	3.9	3.2	3.2	2.9
	No Lid, w/i BL UF	4.3	3.6	3.6	3.3
	Lidded, w/o BL UF	4.2	3.4	3.4	3.2

SOIC

3D packaging is challenging and requires overcoming three major challenges – thermal, power delivery, and yield. The SoIC, as industry-first 3D logic-on-logic and memory-on logic chiplet stacking technology platform, enables the heterogeneous integration (HI) of known good dies (KGDs) with different chip sizes, functionalities, and wafer node technologies, all to be integrated in a single, compact new system chip. From external appearance, SoIC looks like a general SoC chip with multiple pre-designed heterogeneous functional chips embedded. As SoIC is fabricated using "front-end" process, it can be holistically integrated into variant "back-end" advanced packaging technology platforms such as flip chip, integrated fan-out (aka InFO), 3DIC, and 2.5D with Si interposer (e.g. CoWoS) to provide a miniaturized and highly integrated HI SiP for the future HPC, AI, 5G, and edge computing applications. SoIC also offers the design and integration flexibility for mix and match of heterogeneous chips in different technology nodes, materials, functionalities, and chip sizes to create a true heterogeneous 3DIC.

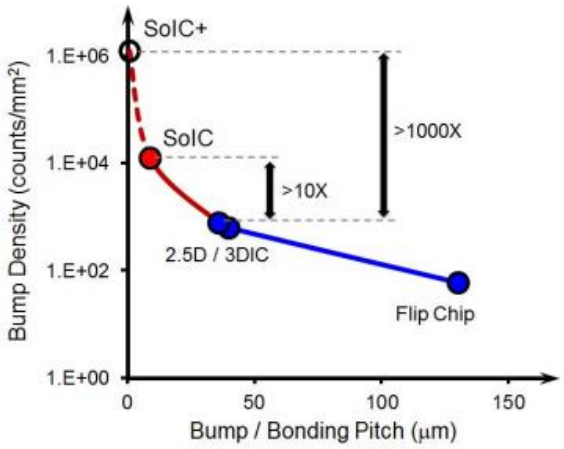
With the closest proximity between two chips through direct chip-to-chip interconnect, the

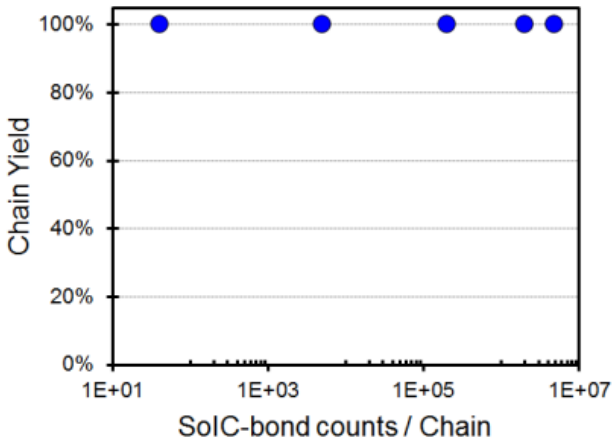




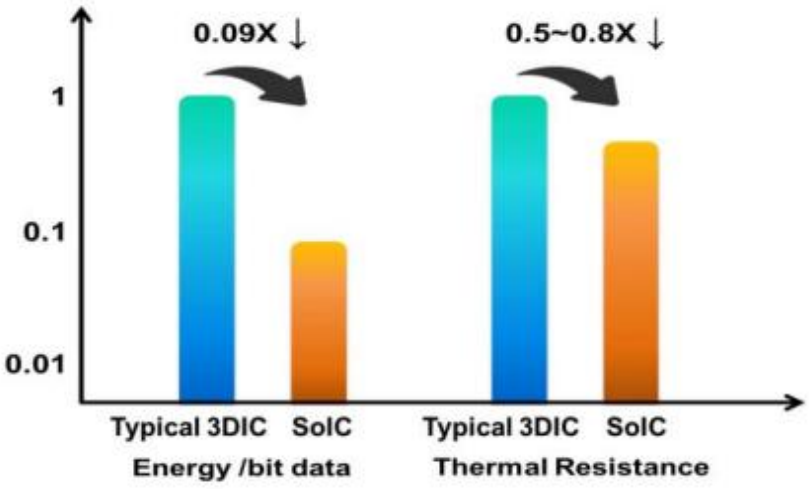
SoIC creates a superior power integrity, signal integrity and a much lower communication latency with more than 20Tbps memory bandwidth to support the future HPC, AI, 5G, and edge computing applications.

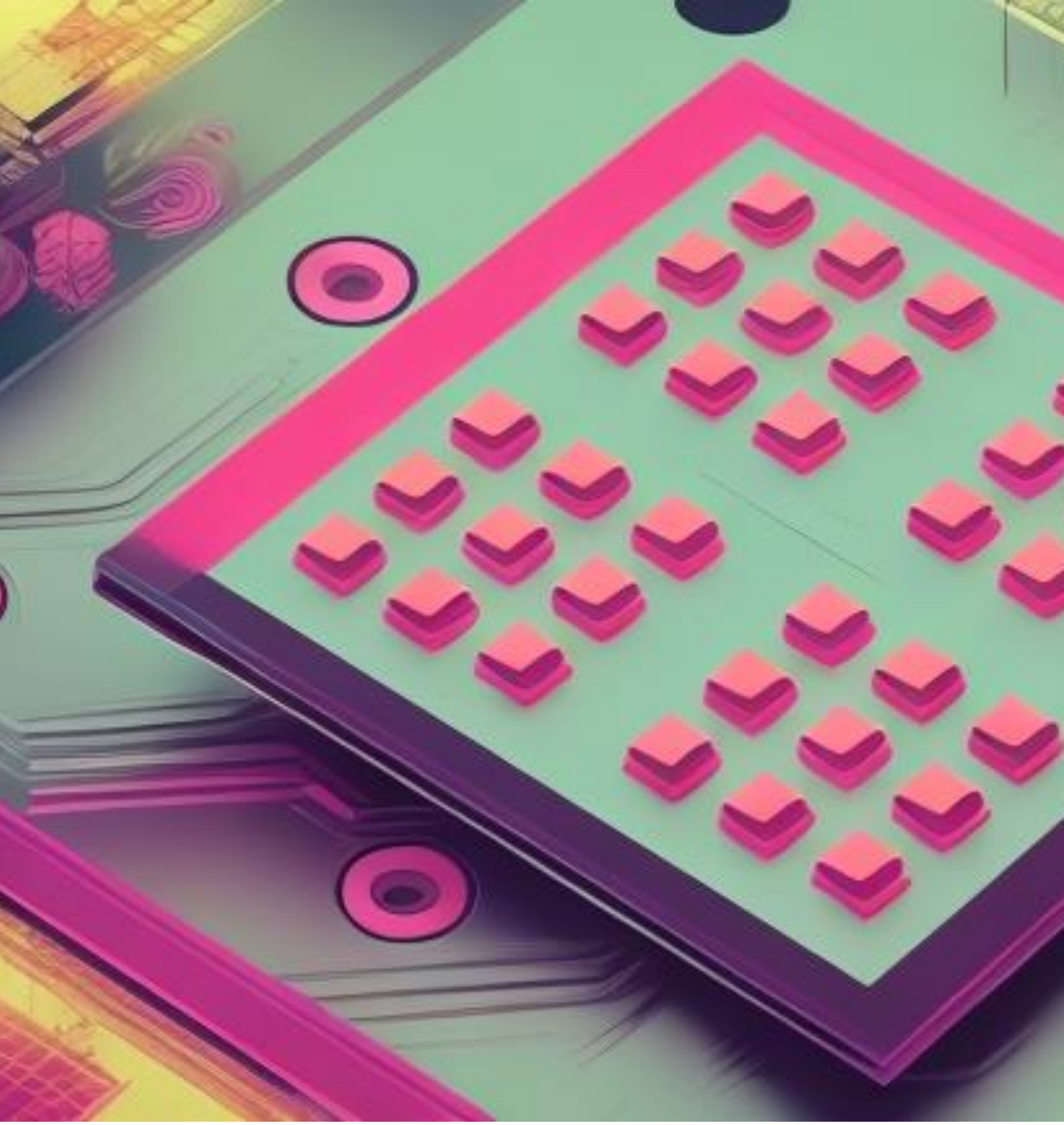
Compared to the typical 3DIC PoP, the SoIC-embedded InFO_PoP offers higher interconnect I/O bonding density, lower power consumption and thinner package profile.the system scaling of bump/bond density can be significantly boosted by SoIC technology to unleash the limitation of flip chip bonding density beyond 10K/mm2. With foundry front-end process control, the integration of SoIC achieves high bonding yield.





SoIC bond outperforms μbump of 3DIC, at 2GHz, by 12.5X in R, by 100X in L, and by 12.5X in C, respectively in normalized scale. As a result, SoIC bond leads μbump bond by 156X in RC delay, and by 12.5X in IR drop. Compared to the typical 3DIC stacking, SoIC by its nature offers higher metal routing density and enables thermal flow in both upward and downward directions , and thus can dissipate the thermal flux more effectively.





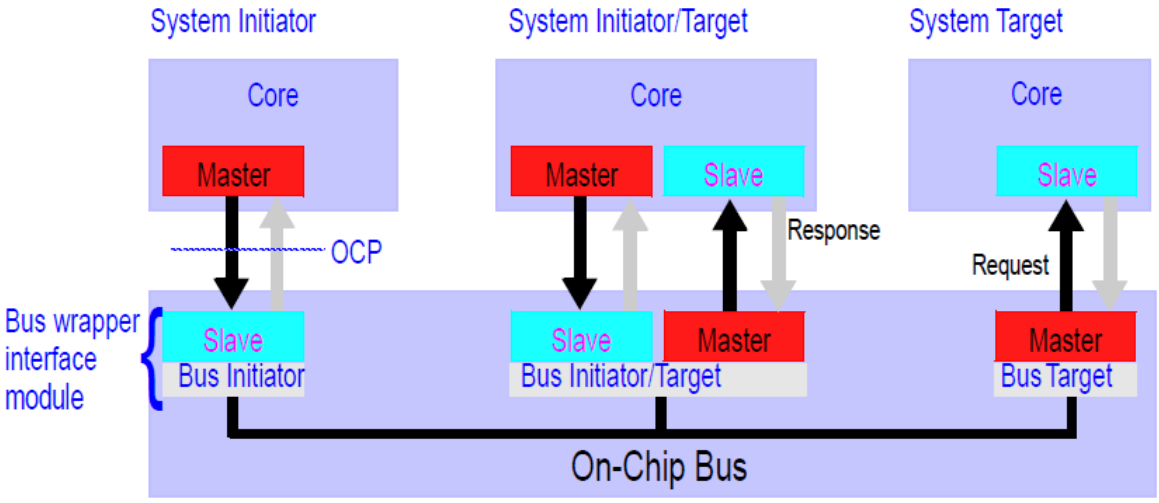
PROTOCOLS & IPs

OPEN CORE PROTOCOL

The Open Core Protocol™ (OCP) defines a high-performance, bus independent interface between IP cores that reduces design time, design risk, and manufacturing costs for SOC designs. Key characteristics of OCP:

- Achieves the goal of IP design reuse. The OCP transforms IP cores, making them independent of the architecture and design of the systems in which they are used.
- Optimizes die area by configuring into the OCP interfaces only those features needed by the communicating cores.
- Simplifies system verification and testing by providing a firm boundary around each IP core that can be observed, controlled, and validated.

The OCP defines a point-to-point interface between two communicating entities, such as IP cores and bus interface modules (bus wrappers). One entity acts as the master of the OCP instance and the other as the slave. Only the master can present commands and is the controlling entity. The slave responds to commands presented to it, either by accepting data from the master, or presenting data to the master. For two entities to communicate in a peer-to-peer fashion, there need to be two instances of the OCP connecting them—one where the first entity is a master, and one where the first entity is a slave.



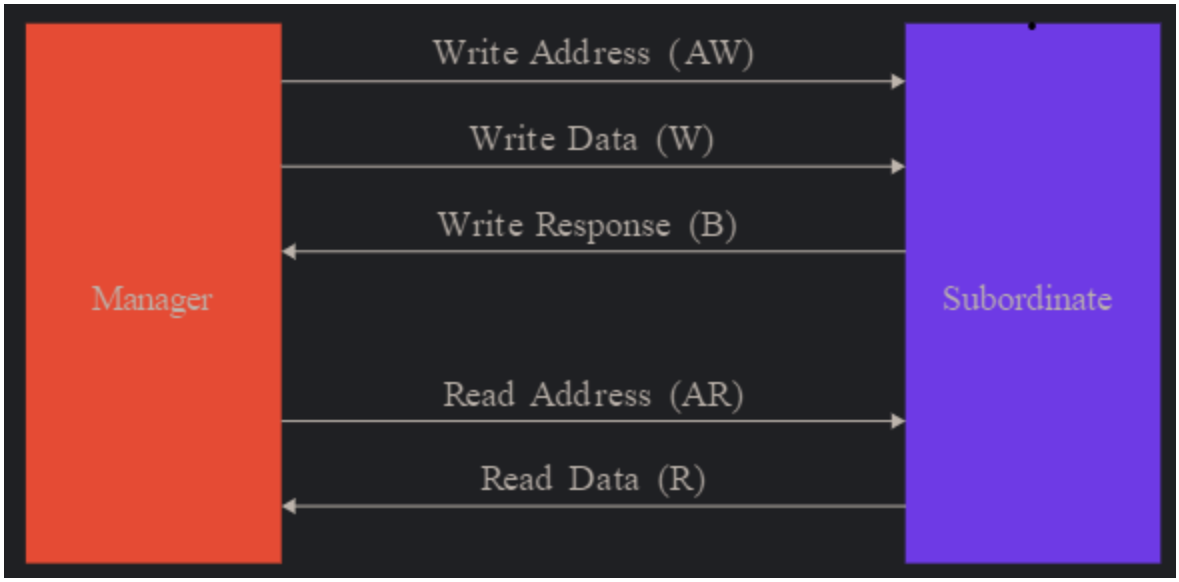
The characteristics of the IP core determine whether the core needs master, slave, or both sides of the OCP. The OCP does not specify the embedded bus functionality. Instead, the interface designer converts the OCP request into an embedded bus transfer. The receiving bus wrapper interface module (as the OCP master) converts the embedded bus operation into a legal OCP command. The system target (OCP slave) receives the command and takes the requested action.

The OCP is flexible. There are several useful models for how existing IP cores communicate with one another. Support for this wide range of behavior is possible with synchronous handshaking signals that allow both the master and slave to control when signals are allowed to change.

AXI

Advanced eXtensible Interface (AXI) is a high-performance interconnect standard that is commonly used in System-on-Chip (SoC) design. It is an open standard that is freely available for use, and it is supported by a wide range of EDA tools and IP vendors.

AXI is a burst-based protocol, meaning that it can transfer multiple data words in a single transaction. This makes it efficient for transferring large amounts of data, such as video or audio streams. AXI also supports several features that make it versatile and easy to use, such as out-of-order transactions, unaligned data transfers, cache support signals, and a low-power interface.



AXI is available in three variants: AXI4, AXI4-Lite, and AXI4-Stream. AXI4 is the most common variant, and it is used for a wide range of applications. AXI4-Lite is a simpler variant that is used for low-power applications. AXI4-Stream is a variant that is specifically designed for high-speed streaming data.

Here are some of the key features of AXI:

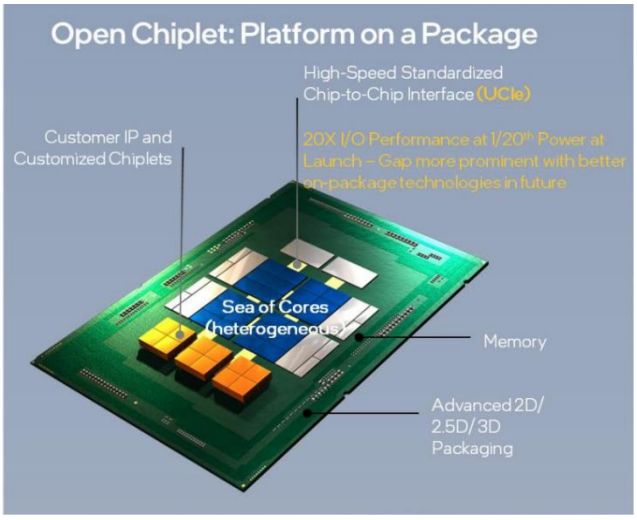
- Burst-based transfers: AXI supports burst-based transfers, which means that it can transfer multiple data words in a single transaction. This makes it efficient for transferring large amounts of data.
- Out-of-order transactions: AXI supports out-of-order transactions, which means that the processor can send multiple requests to the memory controller before waiting for the previous requests to complete. This can improve the performance of the system, especially for applications that have a lot of memory access.
- Unaligned data transfers: AXI supports unaligned data transfers, which means that it can transfer data to any byte address within a memory word. This can be useful for applications that need to access data that is not aligned to the natural boundaries of a memory word.
- Cache support signals: AXI supports cache support signals, which can be used to inform the cache controller about the caching behavior of the memory. This can help to improve the performance of the cache.
- Low-power interface: AXI has a low-power interface, which makes it suitable for use in power-constrained applications.

AXI is a versatile and powerful interconnect standard that is well-suited for a wide range of SoC applications. It is supported by a wide range of EDA tools and IP vendors, and it is freely available for use.

Here are some of the benefits of using AXI:

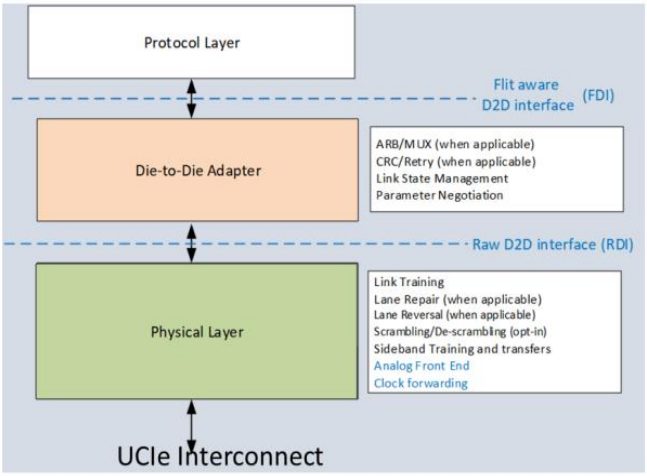
- High performance: AXI can achieve high performance due to its burst-based transfers and out-of-order transactions.
- Flexibility: AXI is a flexible protocol that supports a wide range of features, such as unaligned data transfers and cache support signals.
- Ease of use: AXI is a relatively easy protocol to use, and it is supported by a wide range of EDA tools and IP vendors.
- Free availability: AXI is a freely available open standard.

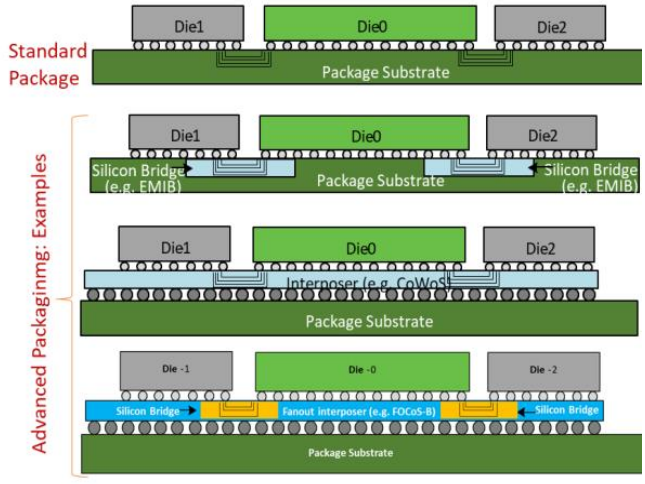
UCIE



The Universal Chiplet Interconnect Express (UCIe) standard serves as an open industry protocol providing a high-bandwidth, low-latency, power-efficient, and cost-effective means for on-package connectivity between chiplets. It caters to the increasing demands across various computing domains, including cloud, edge, enterprise, 5G, automotive, high-performance computing, and handheld devices.

It is a layered protocol designed for chip-to-chip communication in a multi-chip module or system. This protocol is structured with various layers to manage different aspects of data transfer, reliability, compatibility, and interoperability. Here's a breakdown of the key components and functionalities described:





ARCHITECTURE

LAYERS IN UCIE PROTOCOL:

1. Physical Layer:
    - Handles electrical signaling, clocking, link training, and sideband communication.

2. Die-to-Die Adapter:
    - Manages link state, parameter negotiation for chiplets, and ensures reliability through features like cyclic redundancy check (CRC) and link-level retry mechanisms.
    - Defines underlying arbitration mechanisms when multiple protocols are supported.

3. FLIT (Flow Control Unit):
    - Defines the unit of transfer (256-byte FLIT) when the Die-to-Die adapter is responsible for reliable data transfer

PROTOCOL SUPPORT AND INTEROPERABILITY:

NATIVE SUPPORT FOR PCIE AND CXL:

- UCIe natively supports PCIe and CXL protocols, which are widely used across various computing segments.
- This native support ensures seamless interoperability by leveraging the existing ecosystem and infrastructure.

The Universal Chiplet Interconnect Express (UCIe) supports two primary usage models for facilitating efficient and cost-effective performance in the realm of integrated systems:

1. Package Level Integration:

- This model involves integrating various components at the package level. These components can range from memory modules, accelerators, networking devices, modems, and more. This integration can be implemented across a wide spectrum of devices, from handheld gadgets to high-end servers. It allows for connecting dies from multiple sources using various packaging options, even within the same package. This integration aims to deliver power-efficient and cost-effective performance.

2. Off-Package Connectivity:
   - UCIe also facilitates off-package connectivity, enabling connections using various mediums such as optical, electrical cable, or mmWave technology. This is achieved through the utilization of UCIe Retimers, which transport underlying protocols like PCIe and CXL. The objective here is to extend connectivity beyond the node level, allowing for resource pooling, resource sharing, and even message passing using load-store semantics at higher levels such as the rack or pod level. This extension aims to derive better power efficiency and cost-effective performance in edge computing and data centers.

SPECIFICATIONS:

Characteristics / KPIs	Standard Package	Advanced Package	Comments
Characteristics			
Data Rate (GT/s)	4, 8, 12, 16, 24, 32		Lower speeds must be supported -interop (e.g., 4, 8, 12 for 12G device)
Width (each cluster)	16	64	Width degradation in Standard, spare lanes in Advanced
Bump Pitch (um)	100 – 130	25 - 55	Interoperate across bump pitches in each package type across nodes
Channel Reach (mm)	<= 25	<=2	
Target for Key Metrics			
B/W Shoreline (GB/s/mm)	28 – 224	165 – 1317	Conservatively estimated: AP: 45u for AP; Standard: 110u; Proportionate to data rate (4G – 32G)
B/W Density (GB/s/mm^2)	22-125	188-1350	
Power Efficiency target (pJ/b)	0.5	0.25	
Low-power entry/exit	0.5ns <=16G, 0.5-1ns >=24G		Power savings estimated at >= 85%
Latency (Tx + Rx)	< 2ns		Includes D2D Adapter and PHY (FDI to bump and back)
Reliability (FIT)	0 < FIT (Failure In Time) << 1		FIT: #failures in a billion hours (expecting ~1E-10) w/ CXi Flit Mode

1. Variable Parameters:
   - UCIe supports various data rates, widths, bump pitches, and channel reach to ensure the broadest possible interoperability.
2. Cluster Construction:
   - The interconnect is constructed in units called clusters. Each cluster comprises several components, including N single-ended, unidirectional, full-duplex Data Lanes (N = 16 for standard packages and 64 for advanced packages), as well as lanes for Valid, tracking, clock signals, and sideband data
3. **Redundancy and Fault Tolerance:**
   - Advanced packages support spare lanes to handle faulty lanes (clock, valid, sideband, etc.), while standard packages support width degradation to handle failures. This redundancy ensures fault tolerance and system reliability.
4. Scalability:

- Multiple clusters can be aggregated to deliver enhanced performance per link, thereby enhancing scalability and performance.

5. Sideband Interface:
   - UCIe defines a sideband interface, making the design and validation process more straightforward.
   - UCIe aims to cater to diverse integration needs, offering both on-package integration and off-package connectivity options, while ensuring flexibility, fault tolerance, and scalability in its design. This is crucial for achieving efficient and cost-effective performance in a wide range of computing systems.

Utilizing Existing Solutions:

- Leveraging PCIe and CXL enables the use of existing SoC construction, link management, and security solutions in UCIe.
- The usage models addressed include data transfer, memory handling (CXL.Mem), caching requirements for applications (CXL.cache), error handling, and direct memory access, among others.

Innovation and Future Adaptability:

- UCIe defines a "streaming protocol" that can map other protocols.
- The UCIe consortium can innovate on protocols in the future, optimizing them for chiplets as usage models evolve.

Key Objectives:

- Seamless Interoperability: Using PCIe and CXL for compatibility across different computing segments.
- Utilization of Existing Ecosystem: Leveraging the existing infrastructure and solutions for SoC, link management, and security.
- Comprehensive Usage Models: Addressing data transfer, memory, caching, error handling, and more.
- Adaptability for Future Evolution: Allowing for innovation and evolution as usage models change.
- Overall, UCIe aims to provide a standardized, versatile, and adaptable protocol for efficient communication among chiplets within a system, leveraging existing widely-used protocols while allowing for future innovations and advancements.

CCIX

APPLICATION

- Cache Coherent Interconnect for Accelerators or CCIX™ (pronounced 'see 6') is a chip–to-chip interconnect that enables two or more devices to share data in a cache coherent manner. CCIX is poised to optimize and simplify how heterogeneous systems are architected while at the same time increasing bandwidth and reducing latency in the systems built with devices processing via processors with different instruction set architectures (ISAs) or application specific accelerators.
- It is specifically designed for tomorrow's toughest challenges in data centre, cloud computing, Big Data, and any other application where heterogeneous computing is required. The CCIX Standard will be a revolutionary step forward that extends the benefits of open, heterogeneous architecture and cache coherent shared memory model to meet the evolving demand of future data centres.

WORKING PRINCIPLE

- CCIX uses two mechanisms to increase performance and reduce latency. The first mechanism is to use cache coherency to automatically keep the processor and accelerator caches coherent thus facilitating ease-of-use and reducing latency.
- The second mechanism is for CCIX to increase the raw bandwidth of the link. This is done by increasing the maximum link speed to 25GT/s. CCIX Ports can be aggregated together to enable interfaces with performance beyond that of a single interface allowing the accelerator and memory expansion bandwidth to be matched

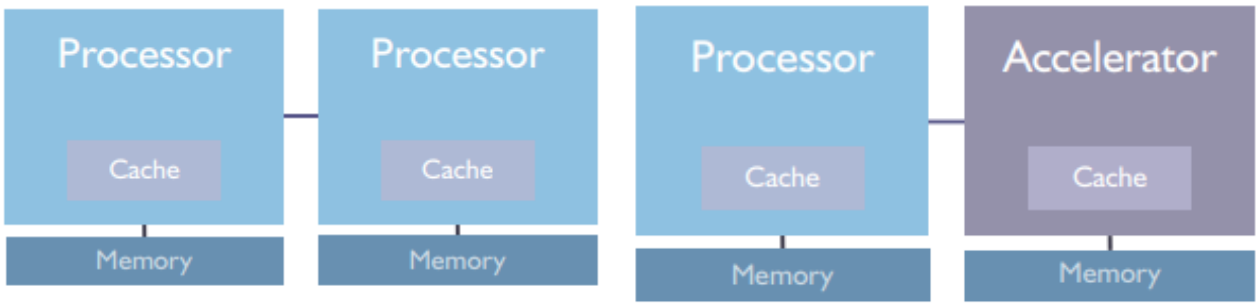


FIGURE 1 | *Multi-processor cache coherency*

- By extending the basic premise of existing cache coherent interconnects to accelerators, application data can be autonomously moved between processor cache and accelerator

cache without a software driver being involved in the data movement. In addition to cache memory, CCIX also enables expansion of the OS paged memory (system memory) to include CCIX device attached memory as well. The CCIX data sharing model is based on use of shared memory addressed with Virtual Addresses (VAs).

- The cache and/or memory between the processor and accelerator is automatically kept up to date using the CCIX protocol. Because the data is automatically synchronized by the CCIX protocol, only data pointers need to be passed rather than relying on complex direct memory access (DMA) drivers. This automatic synchronization leads to reduced data latency and improved application performance. It also simplifies the burden on the software developer allowing them to focus on their application rather than the underlying mechanics of moving data between the accelerator and the host processor.

ARCHITECTURE

- The CCIX architecture is a layer-based architecture that expands on the base PCI Express architecture. CCIX can be thought of as two main specifications that further break down into protocol layers. The CCIX Protocol Specification covers the CCIX Protocol and CCIX Link Layers. These layers define the cache coherent protocol, messaging, and flow control and CCIX transport portions of the protocol. The CCIX Transport Specification, includes the CCIX and PCIe Transaction Layers, the PCIe Data Link Layer and the CCIX Physical Layer. These layers generally take care of the physical link between devices including speed and width negotiation, packet error checking and retry, as well as the initial packet decode protocol.

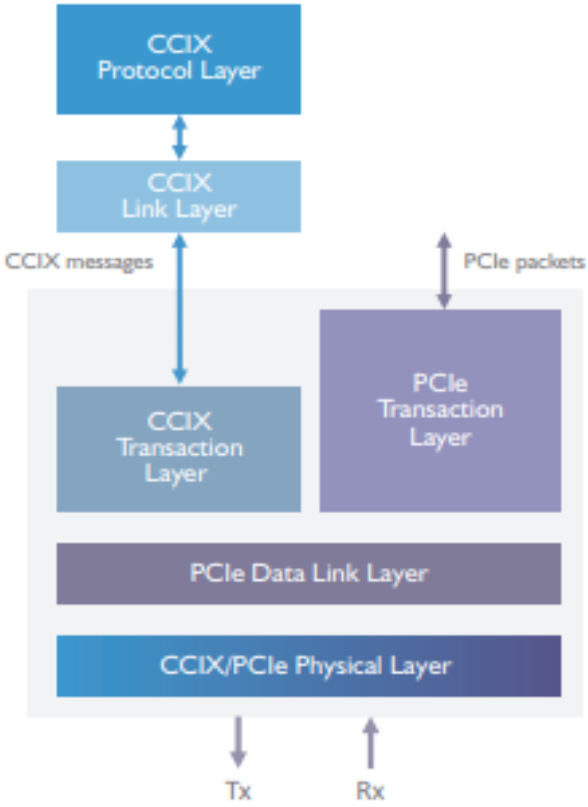


FIGURE 3 | *CCIX Layered Architecture*

CCIX Protocol Layer

- At the very top of the CCIX stack resides the CCIX Protocol Layer. This layer is responsible for the coherency protocol, including memory read and write flows. The layer provides a simple mapping for on-chip coherency protocols such as the Arm AMBA CHI. The cache states defined in this layer allow hardware to determine the state of the memory. For instance, hardware can determine if the data is unique and clean or if it is shared and dirty.

CCIX Link Layer

- The layer below the CCIX Protocol Layer is the CCIX Link Layer. This layer is responsible for formatting CCIX traffic for the target transport. Currently that is PCIe, but with the layered architecture, CCIX could be mapped over a different transport layer in the future. In addition, this layer manages port aggregation, allowing multiple ports to be aggregated together to increase bandwidth.

CCIX and PCIe Transaction Layers

- The CCIX and PCIe transaction layers are responsible for handling their respective packets. The PCIe protocol allows for the implementation of Virtual Channels allowing different data streams to travel across a single PCIe link. By splitting CCIX traffic into one Virtual Channel

and PCIe traffic into a second Virtual Channel, both CCIX and PCIe traffic can share the same link. CCIX can work with either standard PCIe packets or to use optimized CCIX packets. Optimized CCIX packets eliminate several unneeded fields that exist in PCIe packets. When using PCIe packets, existing PCIe switches can be used to transport packets. If optimized CCIX packets are enabled, these optimized packets eliminate the PCIe overhead resulting in a much smaller and more efficient packet for communicating coherency.

PCIe Data Link Layer

- The PCIe Data Link Layer performs all the normal functions of the data link layer. Some examples of these functions are CRC error checking, packet acknowledgment and timeout checking, and credit initialization and exchange.

CCIX/PCIe Physical Layer

- The basis for the CCIX/PCIe Physical layer is the PCIe physical layer. Additionally, CCIX extends the Physical Layer to support 25GT/s. This faster speed, known as ESM (Extended Speed Mode), is automatically detected when two ESM capable devices are connected, resulting in a 56% increase over 16GT/s.

CCIX SYSTEM TOPOLOGY EXAMPLES

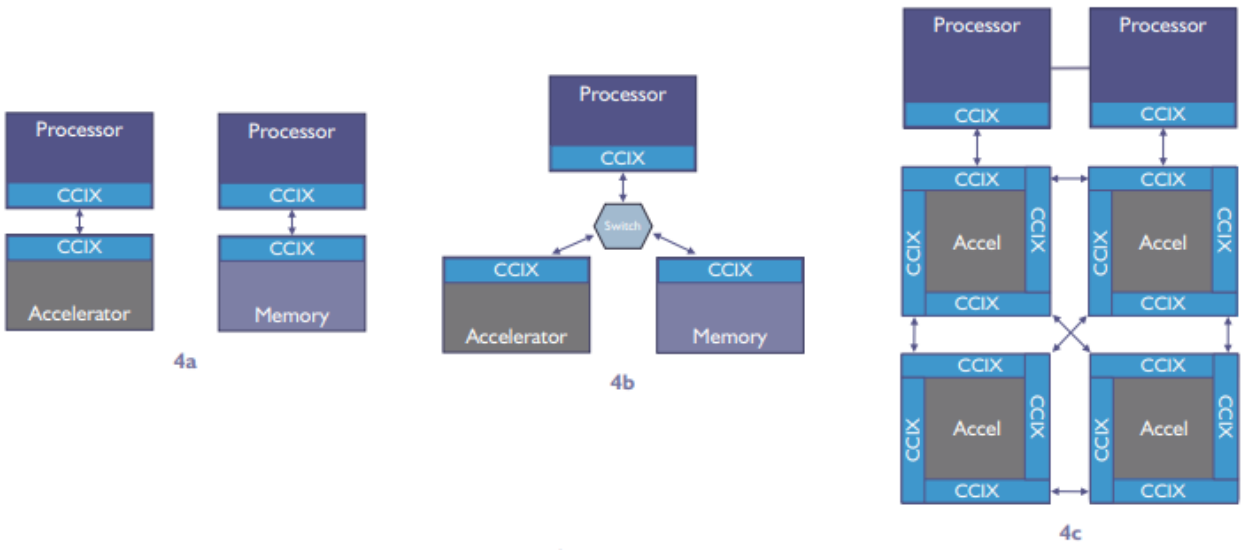


FIGURE 4 | *CCIX Example Topologies*
(4a-Direct Attached; 4b-Switched Topologies; 4c-Hybrid Daisy Chain)

- Because of the layered architecture, CCIX enables several flexible topologies. The most common topology will be direct attached shared virtual memory. But other topologies such as switches, daisy chains and meshes can be easily created and supported.

CCIX DATA FLOWS

Accelerator Shares Processor Memory

- As CCIX is adopted and deployed, the most common initial use case will be for the processor and the accelerator to share host memory and expansion of host memory to include CCIX device attached memory. In this case, there will be two Request Agents, each managing their own caches. The Home Agent will reside in the processor and manages access to the memory that is connected to the processor.

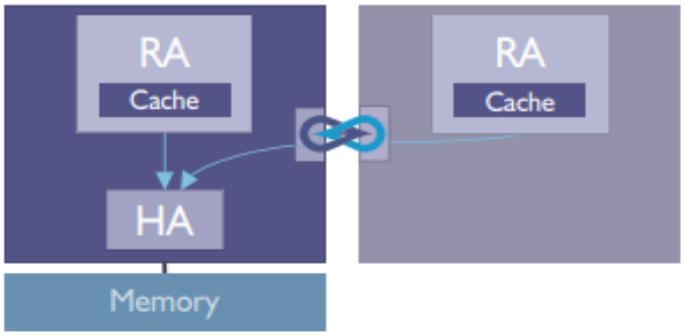


FIGURE 6 | *Accelerator sharing processor memory over CCIX*

Shared Processor and Accelerator Memory

- The next most common model will likely be the case where the processor and accelerator can share virtual memory. In this case, the memory for both the accelerator and the processor are part of a shared virtual memory pool. The processor can simply pass address pointers to the accelerator indicating the data that needs to be worked on, rather than require a complex PCIe DMA and driver to move data between processor and accelerator memory. In this case, there are two Request Agents managing the respective caches and two Home Agents managing memory. By eliminating software driver development and overhead, system performance and software simplicity can be greatly improved.

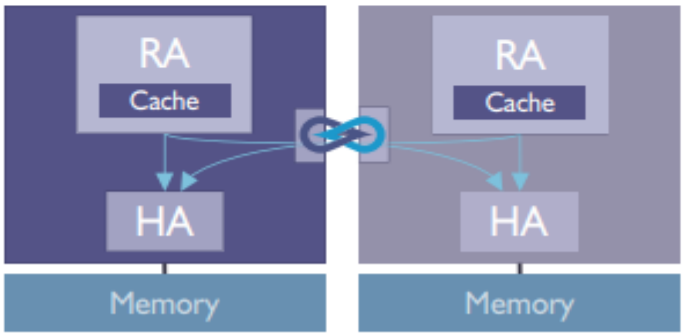


FIGURE 7 | *Shared memory between processor and accelerator*

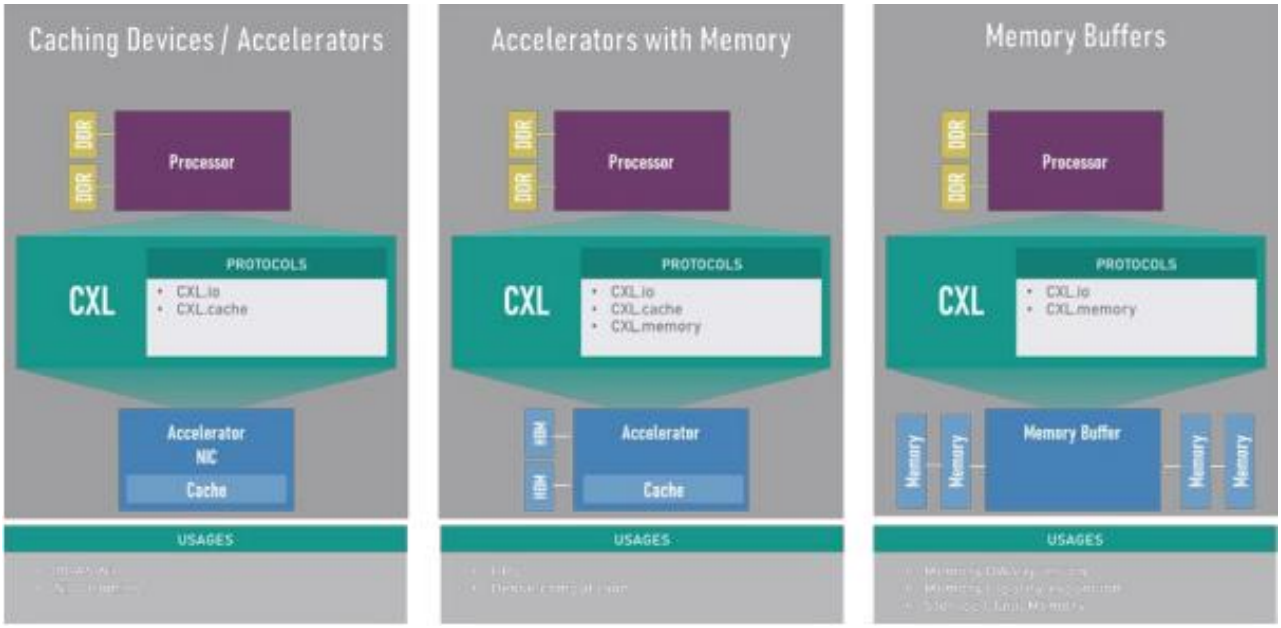
- Beyond the Basic Layout Due to the very flexible nature of CCIX, it can work beyond the basic data flows shown above. From direct attached accelerators to mesh and star networks, CCIX has a very appealing set of options to enable a large variety of topologies.

CXL

Application

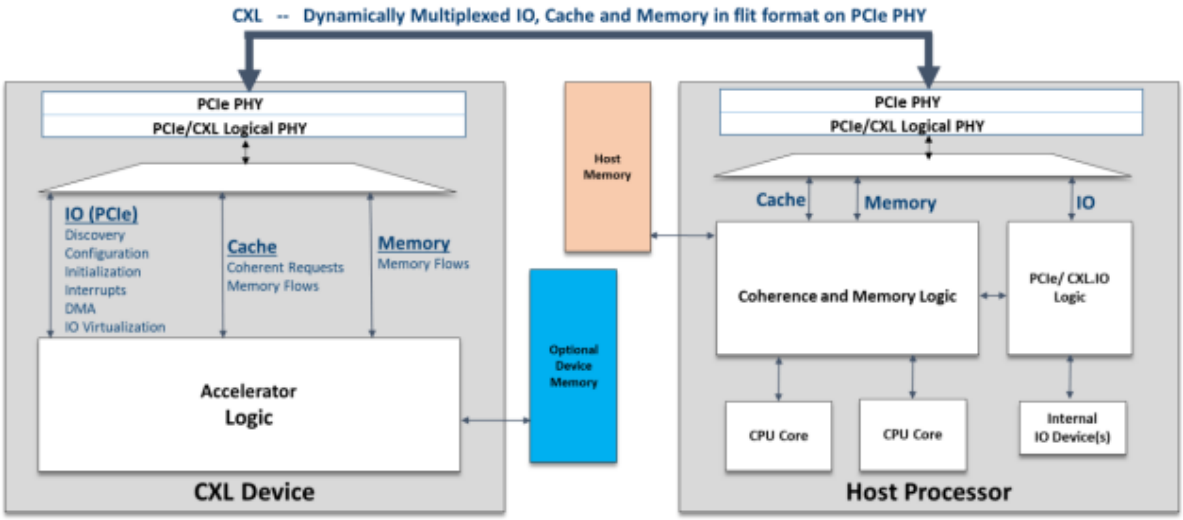
- Compute Express Link™ (CXL™) is an open industry standard interconnect offering high-bandwidth, low latency connectivity between host processor and devices such as accelerators, memory buffers, and smart I/O devices. It is designed to address the growing high-performance computational workloads by supporting heterogeneous processing and memory systems with applications in Artificial Intelligence, Machine Learning, Analytics, Cloud Infrastructure, Cloudification of the Network and Edge, communication systems, and High-Performance Computing by enabling coherency and memory semantics on top of the PCI Express® (PCIe®) based I/O semantics for optimized performance in evolving usage models. This is increasingly important as processing data in these emerging applications requires a diverse mix of scalar, vector, matrix, and spatial architectures deployed in CPU, GPU, FPGA, smart NICs.

ARCHITECTURE



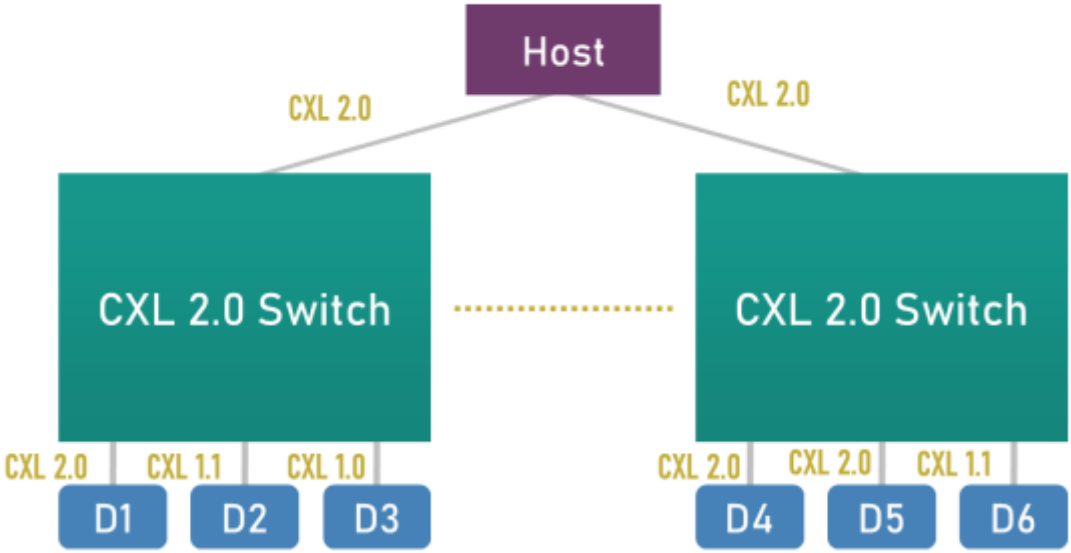
- The CXL.io protocol is based on PCIe and is used for the functions such as device discovery, configuration, initialization, I/O virtualization, and direct memory access (DMA) using noncoherent load-store, producer-consumer semantics. While we expect the PCIe software infrastructure to be reused, device driver would make the necessary enhancements to take advantage of the new capabilities such as CXL.cache and CXL.memory and the system software to program the new set of registers associated with the new capabilities.

- CXL.cache enables a device to cache data from the host memory, employing a simple request and response protocol. The host processor manages coherency of data cached at the device by means of snoop messages.
- CXL.memory allows a host processor to access memory attached to a CXL device. CXL.memory transactions are simple memory load and store transactions that run downstream from the host processor which takes care of all the associated coherency flows.
- By harnessing the capabilities of serial communication and pooling, CXL memory effectively addresses the limitations in performance and socket packaging that are commonly associated with typical DIMM memory solutions. This is particularly advantageous when striving to implement higher storage capacities in computing systems.

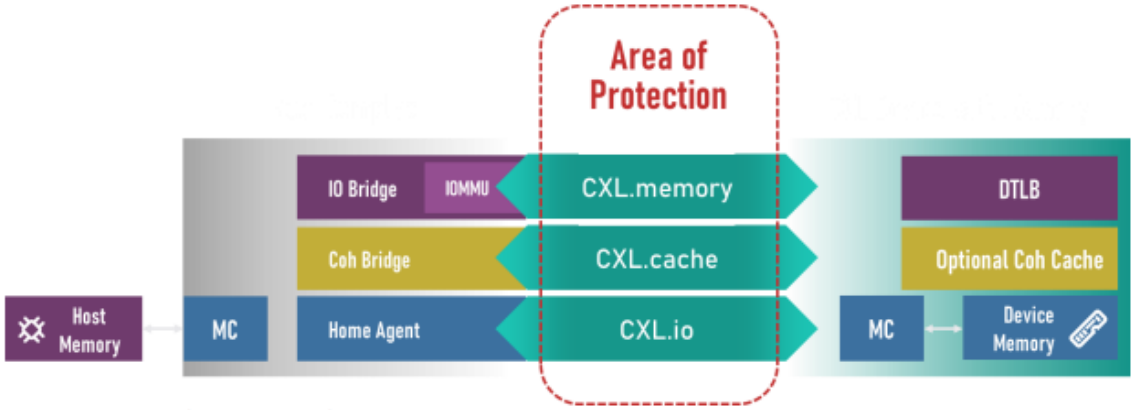


- CXL multiplexes different protocols at the PCIe PHY layer. The unit of transfer for each protocol is flitbased. The CXL.cache and CXL.memory protocols are natively flit based, with CRC protecting each fixedsized flit. CXL.io is packet-based, using the same Transaction Layer Packets (TLPs) and Data Link Layer Packets (DLLPs) as PCIe. The TLP/DLLPs are overlaid on the payload part of the CXL flit. CXL defines policies to deliver the desired quality of service (QoS) across different protocol stacks. The multiplexing of protocols at the PHY level ensures that latency-sensitive protocols such as CXL.cache and CXL.memory have identical low-latency as a native CPU to CPU symmetric coherency link.
- One of new CXL 2.0 features is the support for single level switching to enable fan-out to multiple devices as shown in Figure 2. This will enable many devices in a platform to migrate to CXL, while maintaining the backward compatibility and the low-latency

characteristics of CXL



- One of the important aspects of CXL 2.0 feature set is the support for pooling of multiple logical devices (MLD) as well as single logical device with the help of a CXL switch connected to several Hosts (Root Ports). This feature enables servers to pool resources such as accelerators and/or memory that can be assigned to different servers depending on the workload. Suppose a server needs two FPGAs and a GPGPU, it can ask for those resources from the resource manager in the rack and obtain those if available and relinquish them when its job is done.



- Security is a key cornerstone for any technology to be successful, considering the vulnerability attacks that are so pervasive. CXL is making great strides in this regard, working collaboratively with other industry-standard bodies such as PCI-SIG and DMTF to ensure that we have a seamless user experience while providing the best security

mechanisms. CXL 2.0 enables encryption on the Link that works seamlessly with existing security mechanisms such as device TLB.

- CXL 3.0, which is based on PCIe 6.0 technology, doubles the transfer rate to 64GT/s with no additional latency over previous generations. This allows for aggregate raw bandwidth of up to 256GB/s for x16 width link. For low-latency transfers, CXL 3.0 leverages PCIe 6.0's combination of lightweight FEC and strong CRC for error free transmission with 256B flits on PAM-4 signalling to achieve 64GT/s. CXL 3.0, however, goes further in introducing a latency-optimized flit variant to further reduce 2-5ns of latency by breaking up the CRC in 128B sub-flit granular transfers to mitigate store-and-forward overheads in the physical layer. As with previous generations of CXL, the new 256B flit format is backward compatible with 8GT/s, 16GT/s and 32GT/s which eases the transition to CXL 3.0.
- A major enhancement in CXL 3.0 is the ability to back invalidate the Host's caches. This model of maintaining coherency for host-managed device-attached memory (HDM) is called enhanced coherency and replaces Bias Based coherency introduced in previous generations.
- With CXL 3.0, in addition to memory pooling, the concept of memory sharing is introduced. Memory sharing is the ability of CXL-attached memory to be coherently shared across hosts using hardware coherency. Thus, unlike memory pooling, memory sharing allows the a given region of memory to be simultaneously accessible by more than one host and still guarantee that every host sees the most up to date data at that location, without the need for software-managed coordination. This allows system designs to build clusters of machines to solve large problems through shared memory constructs.
- CXL 3.0 introduces fabric capabilities which goes beyond traditional tree based architectural structures of PCIe and previous CXL generation. The CXL fabric can support up to 4096 nodes that can communicate with each other using a new scalable addressing mechanism called Port Based Routing (PBR). Here, a node can be a CPU Host, a CXL accelerator with or without memory, a PCIe device or a Global Fabric Attached Memory (GFAM) device. A GFAM device is like a traditional CXL Type-3 device, except it can be accessed by multiple nodes (up to 4095) in flexible ways using port-based routing. This architecture opens a world of possibilities in constructing powerful systems comprising of compute and memory elements arranged to suffice the needs of workloads.

SPECIFICATION

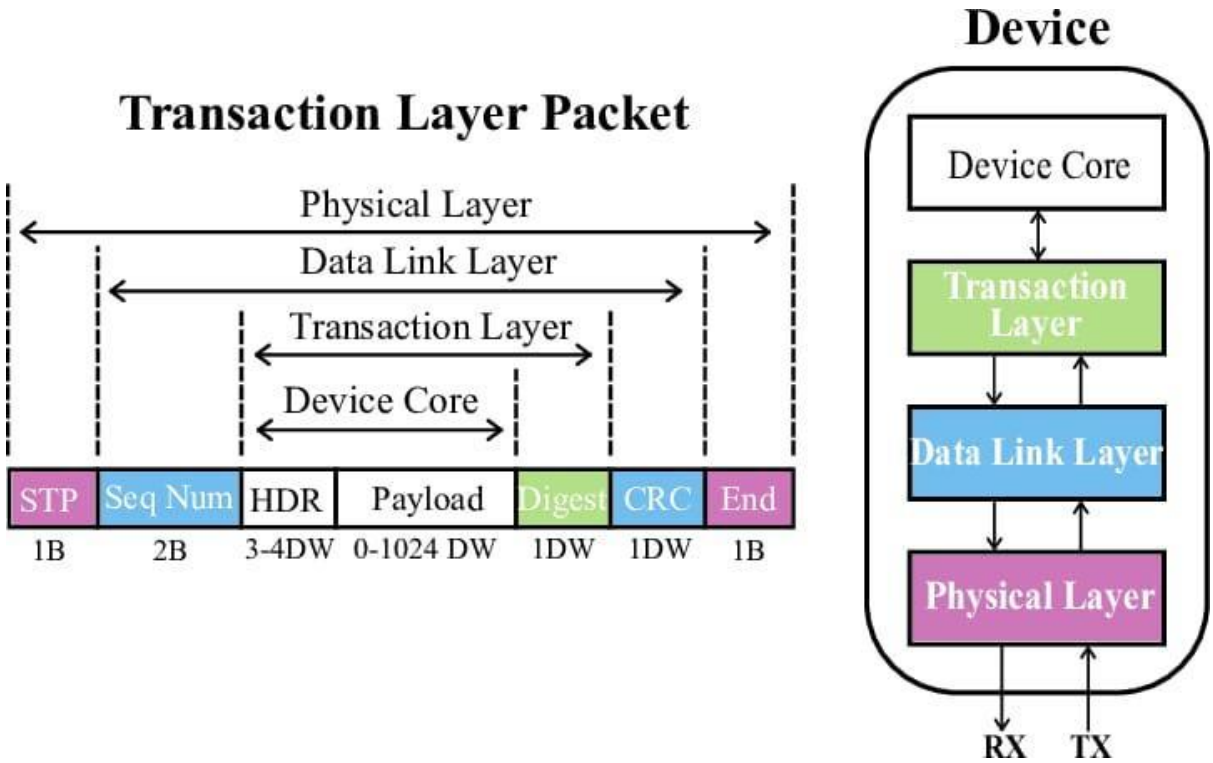
Features	CXL 1.0 / 1.1	CXL 2.0	CXL 3.0
Release date	2019	2020	1H 2022
Max link rate	32GTs	32GTs	64GTs
Flit 68 byte (up to 32 GTs)	✓	✓	✓
Flit 256 byte (up to 64 GTs)			✓
Type 1, Type 2 and Type 3 Devices	✓	✓	✓
Memory Pooling w/ MLDs		✓	✓
Global Persistent Flush		✓	✓
CXL IDE		✓	✓
Switching (Single-level)		✓	✓
Switching (Multi-level)			✓
Direct memory access for peer-to-peer			✓
Enhanced coherency (256 byte flit)			✓
Memory sharing (256 byte flit)			✓
Multiple Type 1/Type 2 devices per root port			✓
Fabric capabilities (256 byte flit)			✓

PCIE

Application

- PCI Express (Peripheral Component Interconnect Express), often abbreviated as PCIe or PCI-e, is a high-speed serial computer expansion bus standard. It was designed to supplant the aging PCI (Peripheral Component Interconnect), PCI-X, and AGP (Accelerated Graphics Port) bus standards. PCIe serves as the primary motherboard interface for various components in personal computers, such as graphics cards, sound cards, host adapters for hard disk drives, Solid State Drives (SSDs), Wi-Fi and Ethernet hardware connections.
- Apart from its primary role in personal computers, the PCIe interface is employed in various other standards. Notably, it is the foundation for the laptop expansion card interface called ExpressCard. Furthermore, it is utilized in storage interfaces such as SATA Express, U.2 (SFF-8639), and M.2, contributing to high-speed and efficient data transfer between storage devices and the computer.

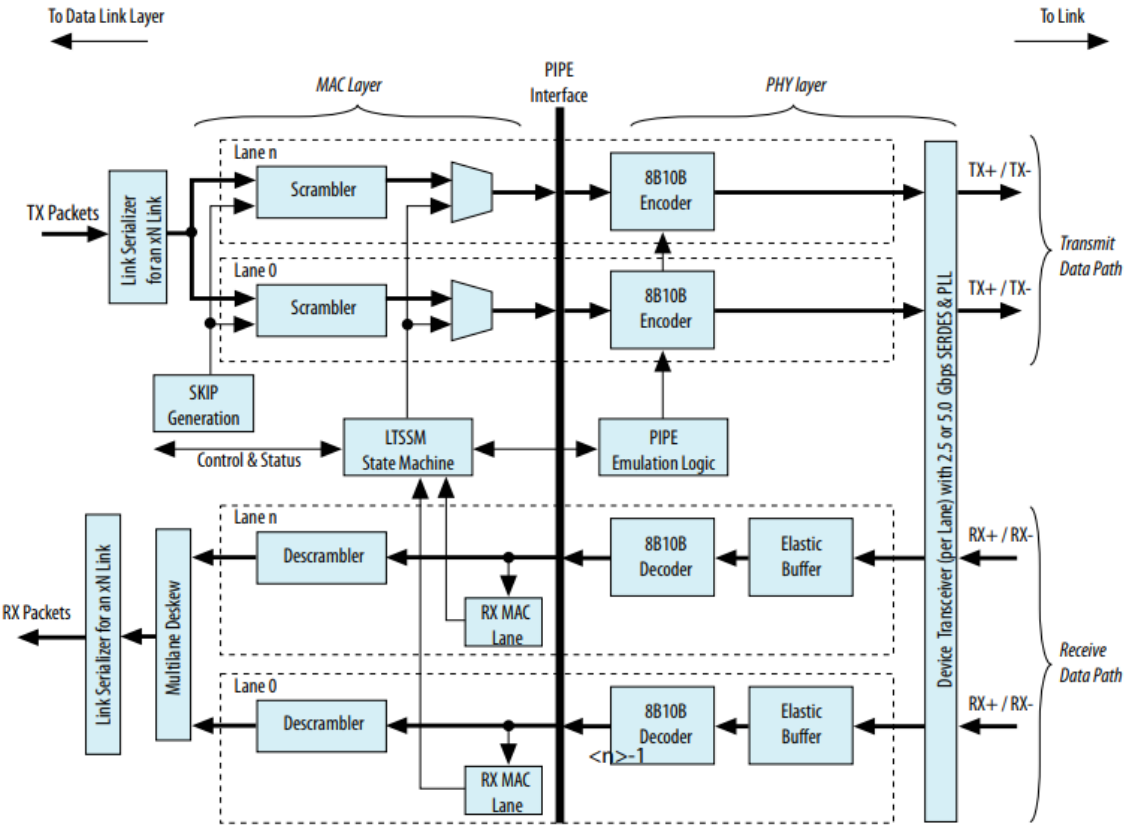
ARCHITECTURE



- PCIe implements a point-to-point topology with separate serial links connecting each device directly to the root complex (host).

- PCIe communicates using packets, which are handled by the transaction layer. This packet-based communication is a radical departure from the older scheme of electrical signalling and protocol.
- The PCI Express (PCIe) link utilizes dedicated unidirectional serial connections known as "lanes" as opposed to the shared bidirectional parallel bus system used in earlier PCI connections.

Physical Layer Architecture



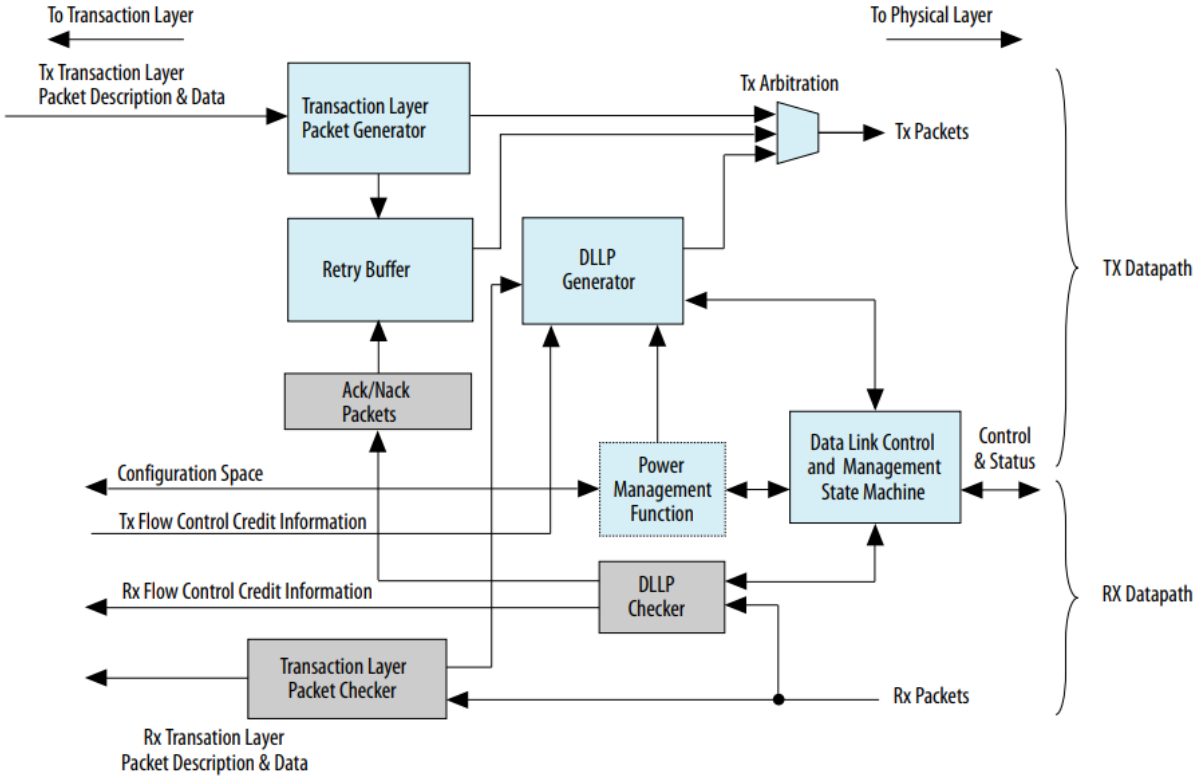
Physical Layer:

- Lanes and Connectors: PCIe employs serial connections known as lanes, each consisting of two unidirectional differential pairs with one pair for receiving data and the other for transmitting. Thus, each lane is composed of four wires or signal traces. Conceptually, each lane is used as a full-duplex byte stream, transporting data packets in eight-bit "byte" format simultaneously in both directions between endpoints of a link. The physical layer of PCIe uses various lane configurations, ranging from x1 to x16, with different numbers of pins and lengths for connectors corresponding to the lanes. These lanes can operate at different speeds, such as 2.5, 5, 8, 16, or 32 Gbit/s, depending on negotiated capabilities.

Data Transmission:

- Data transmission over multiple-lane links is interleaved. This means that data is sent down successive lanes, which, although requiring hardware complexity to synchronize, can reduce latency. PCIe employs encoding schemes to maintain synchronization and data integrity. For instance, PCIe 2.0 uses 8b/10b encoding, while PCIe 3.0 uses 128b/130b encoding to improve available bandwidth and ensure synchronization.

Data Link Layer



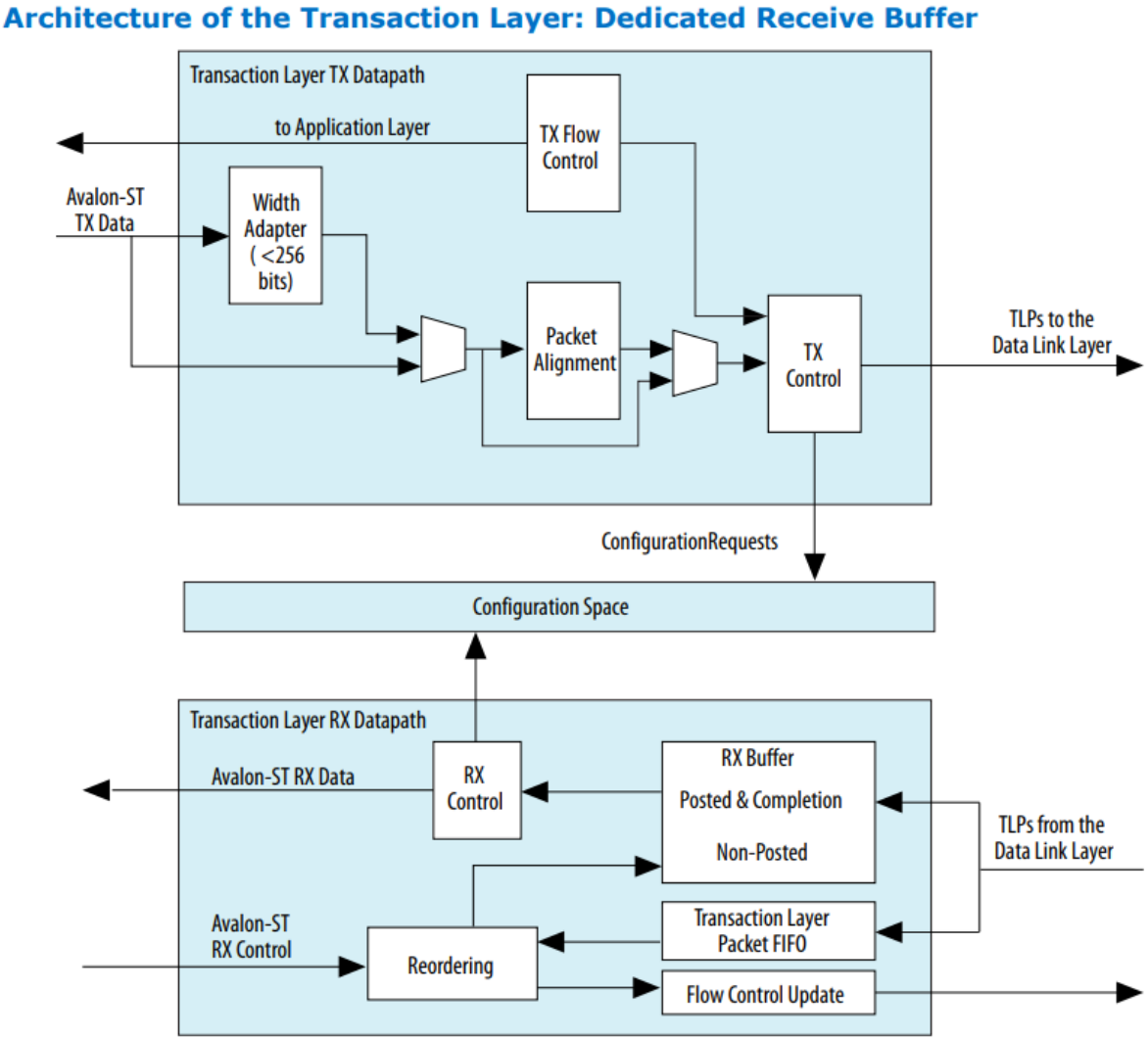
Data Link Layer:

- The data link layer handles services such as sequencing transaction layer packets(TLPs), ensuring reliable delivery via an acknowledgement protocol, and managing flow control credits.

1. Sequence the transaction layer packets (TLPs) that are generated by the transaction layer,
2. Ensure reliable delivery of TLPs between two endpoints via an acknowledgement protocol (ACK and NAK signaling) that explicitly requires replay of unacknowledged/bad TLPs,
3. Initialize and manage flow control credits

- Sequence Numbering and Error Handling: Each outgoing packet is given an incrementing sequence number, and a 32-bit cyclic redundancy check (CRC) code is added. The link layer validates the received packet using these details and, if invalid, requests re-transmission. The receiver sends negative acknowledgments (NAK) for faulty packets, prompting re-transmission by the sender. ACK messages are sent for successful packet reception.

FLOW CONTROL AND RELIABILITY:



- In-Flight TLPs and Credits: The number of unacknowledged TLPs on the link is constrained by the transmitter's replay buffer and flow control credits issued by the receiver to the transmitter. This mechanism regulates the flow of data and credits to maintain stable communication.
- PCIe's design, with dedicated serial lanes and a structured layered protocol, allows for high-speed, reliable, and efficient data transfer, ensuring compatibility and negotiating the best link configurations between devices. Serial interfaces offer advantages such as higher bandwidth, reduced timing skew issues, duplex communication, and greater flexibility in lane allocation for different devices, making them a more efficient and scalable choice for modern computing and digital communication standards. The protocol's robust error-

handling mechanisms and flow control contribute to its overall reliability in transmitting data across various devices.

SPECIFICATIONS

PCI Express link performance[47][48]

Version	Introduced	Line code		Transfer rate per lane[i][ii]	Throughput[i][iii]				
					x1	x2	x4	x8	x16
1.0	2003	NRZ	8b/10b	2.5 GT/s	0.250 GB/s	0.500 GB/s	1.000 GB/s	2.000 GB/s	4.000 GB/s
2.0	2007			5.0 GT/s	0.500 GB/s	1.000 GB/s	2.000 GB/s	4.000 GB/s	8.000 GB/s
3.0	2010		128b/130b	8.0 GT/s	0.985 GB/s	1.969 GB/s	3.938 GB/s	7.877 GB/s	15.754 GB/s
4.0	2017			16.0 GT/s	1.969 GB/s	3.938 GB/s	7.877 GB/s	15.754 GB/s	31.508 GB/s
5.0	2019			32.0 GT/s	3.938 GB/s	7.877 GB/s	15.754 GB/s	31.508 GB/s	63.015 GB/s
6.0	2022	PAM-4 FEC	1b/1b 242B/256B FLIT	64.0 GT/s 32.0 GBd	7.563 GB/s	15.125 GB/s	30.250 GB/s	60.500 GB/s	121.000 GB/s
7.0	2025 (planned)			128.0 GT/s 64.0 GBd	15.125 GB/s	30.250 GB/s	60.500 GB/s	121.000 GB/s	242.000 GB/s

Notes

i. ^ [a] [b] In each direction (each lane is a dual simplex channel).

ii. ^ Transfer rate refers to the encoded serial bit rate; 2.5 GT/s means 2.5 Gbit/s serial data rate.

iii. ^ Throughput indicates the unencoded bandwidth (without 8b/10b, 128b/130b, or 242B/256B encoding overhead). The PCIe 1.0 transfer rate of 2.5 GT/s per lane means a 2.5 Gbit/s serial bit rate corresponding to a throughput of 2.0 Gbit/s or 250 MB/s prior to 8b/10b encoding.

PACKET EFFICIENCY CALCULATION:

Gen 2 Transaction Layer Packet[119]:3

Layer	PHY	Data Link Layer	Transaction			Data Link Layer	PHY
Data	Start	Sequence	Header	Payload	ECRC	LCRC	End
Size (Bytes)	1	2	12 or 16	0 to 4096	4 (optional)	4	1

The Gen2 overhead is then 20, 24, or 28 bytes per transaction.[*clarification needed*][*citation needed*]

Gen 3 Transaction Layer Packet[119]:3

Layer	G3 PHY	Data Link Layer	Transaction Layer			Data Link Layer
Data	Start	Sequence	Header	Payload	ECRC	LCRC
Size (Bytes)	4	2	12 or 16	0 to 4096	4 (optional)	4

The Gen3 overhead is then 22, 26 or 30 bytes per transaction.[*clarification needed*][*citation needed*]

Gen 2 Transaction Layer Packet[119]:3

Layer	PHY	Data Link Layer	Transaction			Data Link Layer	PHY
Data	Start	Sequence	Header	Payload	ECRC	LCRC	End
Size (Bytes)	1	2	12 or 16	0 to 4096	4 (optional)	4	1

The Gen2 overhead is then 20, 24, or 28 bytes per transaction.[*clarification needed*][*citation needed*]

Gen 3 Transaction Layer Packet[119]:3

Layer	G3 PHY	Data Link Layer	Transaction Layer			Data Link Layer
Data	Start	Sequence	Header	Payload	ECRC	LCRC
Size (Bytes)	4	2	12 or 16	0 to 4096	4 (optional)	4

The Gen3 overhead is then 22, 26 or 30 bytes per transaction.[*clarification needed*][*citation needed*]

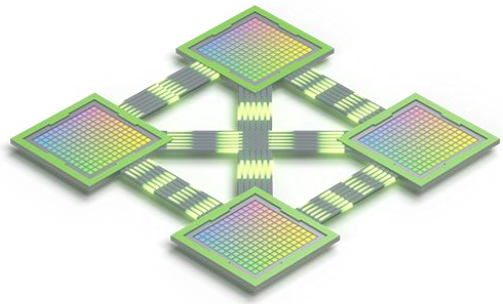
The packet efficiency, which is a measure of how much of the total transmitted data is actual payload, can be calculated using the formula:

$$\text{Packet Efficiency} = \frac{\text{Payload}}{\text{Payload} + \text{Overhead}}$$

- For a 128-byte payload, the packet efficiency is around 86%.
- For a larger 1024-byte payload, the packet efficiency increases to about 98%.

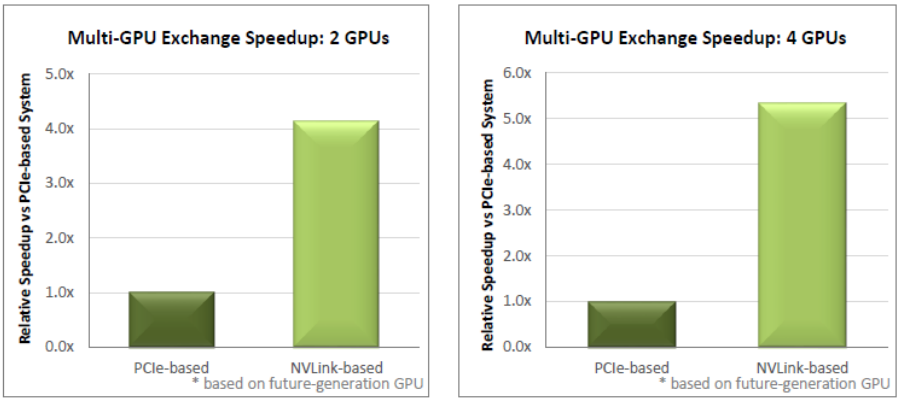
NVLink:

NVLink is a high-speed interconnect technology developed by Nvidia, designed to ease swift data exchange and computation collaboration between GPU and CPU processors in accelerated systems. Unlike PCI Express, NVLink is a wire-based serial multi-lane near-range communications link that supports multiple connections within a single device. Devices equipped with NVLink utilize a mesh networking approach for communication rather than relying on a central hub. This technology empowers processors to transmit and receive data from shared memory pools at remarkable speeds, accelerating data processing. Additionally, NVLink incorporates resiliency features like link-level error detection and packet replay mechanisms, ensuring the reliable and error-free transmission of data.



NVIDIA H100 with NVLink GPU-to-GPU connections

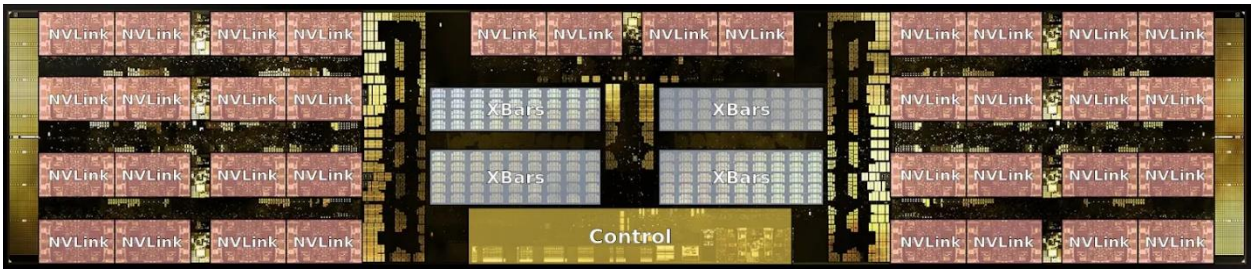
NVLink leverages Nvidia's High-Speed Signaling interconnect (NVHS), a bidirectional interface. Each NVLink connection has eight differential pairs in each direction, totaling 32 wires. Data is transmitted up to 20 Gbps per second, resulting in a peak bidirectional bandwidth of 40 Gbps for a single NVLink connection. The latest 4th generation of NVLink achieves an impressive bandwidth of 900GB/s. In Nvidia Hopper architecture shown above, for instance, each GPU has 18 NVLink links, further enhancing data transfer capabilities. NVLink's performance surpasses traditional PCIe interfaces, enabling faster and more efficient data communication between GPUs and CPUs. This can be seen from the analysis for 2 and 4 GPU exchange with communication based on NVLINK and PCLE.



Projected multi-GPU exchange performance in 2-GPU and 4-GPU configurations, comparing NVLink-based systems to PCIe-based systems

NVSwitch:

NVLink has significantly enhanced the communication speed between GPUs in multi-GPU systems. This technology allows for the design of such systems using multiple NVLink connections. NVSwitch, on the other hand, is an 18-port NVLink switch developed by Nvidia. It is built on TSMC's 12 nm process and comprises 2 billion transistors. The NVSwitch offers an impressive total bandwidth of 900 GB/s, making it a powerful interconnect solution for high-performance computing. As of now, NVSwitch is primarily employed in Nvidia's own DGX-2 AI computer, where it plays a crucial role in facilitating efficient communication and data sharing among multiple GPUs. It serves as a principal component to maximize the performance of AI and high-performance computing workloads on the DGX-2 platform.



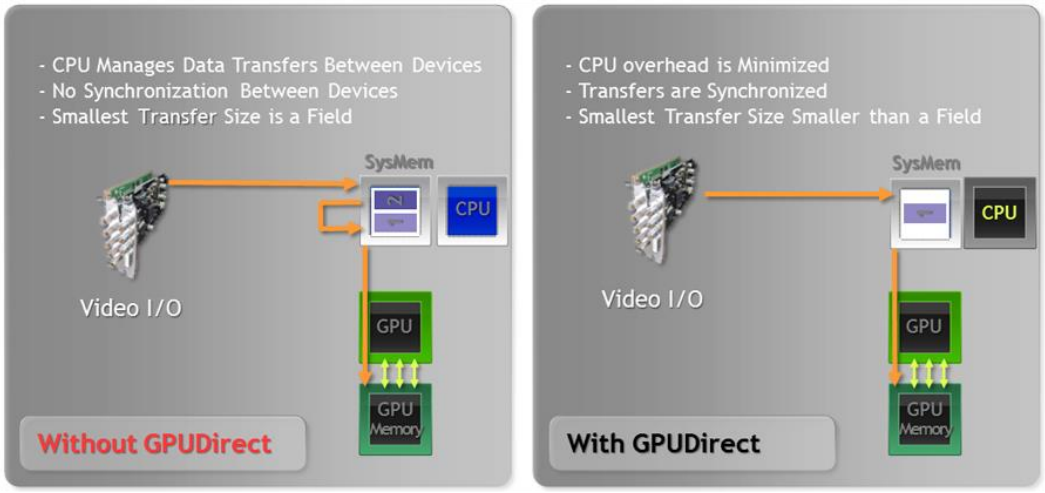
NVSwitch chip die

NVSwitch chips operate in parallel to enable the interconnection of increasingly more significant numbers of GPUs in a highly scalable manner. In a closed system with eight GPUs, for instance, three NVSwitch chips can be utilized. Each GPU connects to each switch through two NVLink paths, ensuring robust communication between all GPUs in the system.

The design of these NVSwitch chips allows for the interleaving of data traffic across all the NVLinks and NVSwitches, effectively optimizing data transfer and reducing bottlenecks. This means that GPUs can communicate with each other pairwise, utilizing the entire 300 Gbps bidirectional bandwidth available between any pair of GPUs. The NVSwitch chips supply unique and efficient paths from any source GPU to any destination GPU, ensuring high-speed and reliable data exchange in multi-GPU systems.

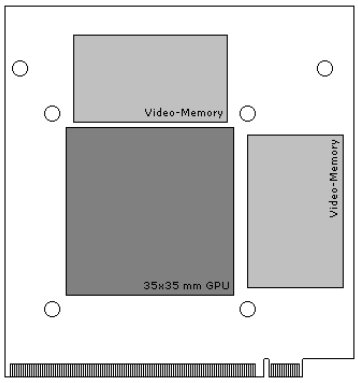
GPU Direct:

GPUDirect is a technology developed by NVIDIA to streamline and enhance data transfers within a computing system, particularly between GPUs and other hardware components. Its primary purpose is to minimize data transfer latency and boost overall application performance by enabling GPUs to access data directly, eliminating the need for redundant data copying through CPU memory. GPUDirect comes in various iterations, such as GPUDirect RDMA for efficient network communication and GPUDirect Storage for optimized access to storage data. These technologies collectively facilitate faster and more efficient data transfer to and from GPUs, making them exceptionally valuable for high-performance computing, artificial intelligence, and data-intensive applications. When combined with NVLink technology, it can further enhance data throughput and reduce latency, improving system performance and responsiveness.



These technologies can revolutionize the automotive industry by facilitating more efficient data processing and communication between GPUs and other components. This is a critical aspect for the development and operation of autonomous vehicles. They are particularly beneficial for autonomous driving systems, where there is a need to process large volumes of sensor data in real-time. The ability to process and analyze data quickly is vital for real-time decision-making in autonomous vehicles. These advancements contribute significantly to the safety and efficiency of the vehicles as well.

MOBILE PCI EXPRESS MODULE

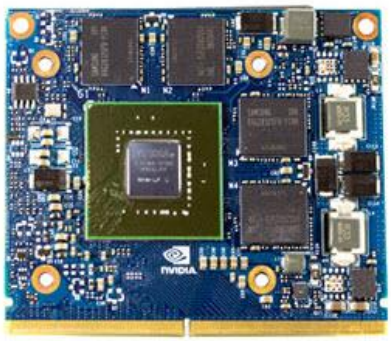


Design of a first-generation MXM-II card for 35 mm GPU

Mobile PCI Express Module (MXM) is a standardized GPU interconnect technology designed specifically for laptops, using the high-speed PCI Express interface. MXM-SIG developed MXM and serves as a common platform for connecting essential components like graphics units and video cards within laptops, ensuring seamless communication between them. Moreover, MXM modules offer compatibility with GPUs from multiple manufacturers, including NVIDIA and AMD.

There are two MXM modules, available in third generation, Type A and Type B. Each class has a distinct form factor targeted for different performance, power, and thermal requirements.

MXM Type	Width	Length	Max. Power	GPU Memory Bus
MXM-A	82mm	70mm	55W	64-bit or 128-bit
MXM-B	82mm	105mm	200W	256-bit



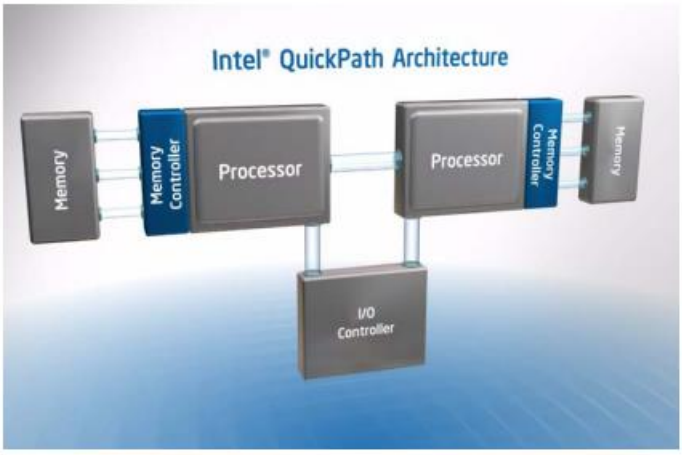
MXM Type A

MXM Type B

MXM generation 3 supports the following key features:

- ❑ Up to 16 lanes PCI Express V3.0
- ❑ Up to 8 DDR2, DDR3, GDDR3 or GDDR5 memory devices
- ❑ Up to six Dual-mode DisplayPort (with support for EDP, DVI and HDMI)
- ❑ Single 24-bit dual-link LVDS, dual-link DVI and HDMI
- ❑ Single VGA and TV-out.

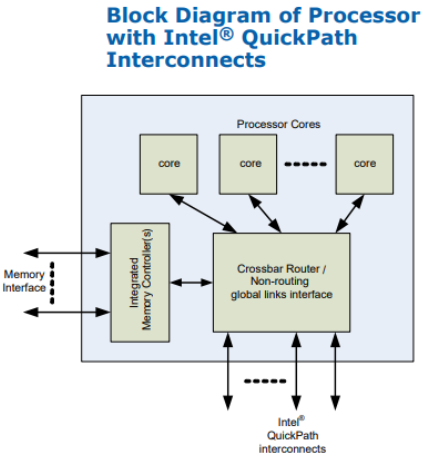
INTEL QUICKPATH INTERCONNECT

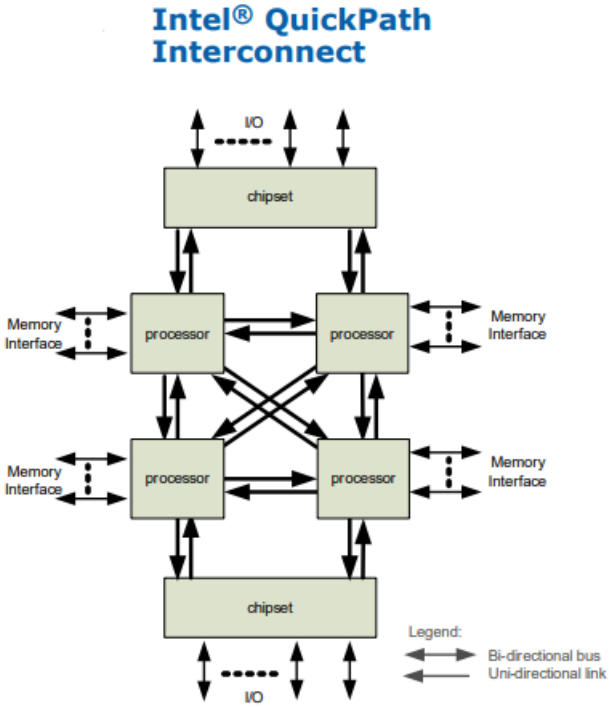


The Intel® QuickPath Interconnect is a highspeed, packetized, point-to-point interconnect used in Intel's next generation of microprocessors first produced in the second half of 2008. The narrow high-speed links stitch together processors in a distributed shared memory1-style platform architecture. Compared with today's wide front-side buses, it offers much higher bandwidth with low latency. The Intel® QuickPath Interconnect has an efficient architecture allowing more interconnect performance to be achieved in real systems. It has a snoop protocol optimized for low latency and high scalability, as well as packet and lane structures enabling quick completions of transactions. Reliability, availability, and serviceability features (RAS) are built into the architecture to meet the needs of even the most mission-critical servers. With this compelling mix of performance and features, it's evident that the Intel® QuickPath Interconnect provides the foundation for future generations of Intel microprocessors and that various vendors are designing innovative products around this interconnect technology.

1.Structure of Interconnect

The Intel® QuickPath Interconnect (QPI) serves as a high-speed, point-to-point processor interconnect, designed for high-bandwidth communication among Intel processors in a variety of applications. It's divided into several layers: Link, Routing, Transport, Protocol, and Agents.





(a) Link Layer

- The Link layer manages the physical connection, error detection, and recovery features of the interconnect.
- Features like self-healing links and clock fail-over ensure error recovery without software intervention.
- Unrecoverable soft errors prompt dynamic link width reduction, allowing negotiation between sender and receiver to connect through reduced link widths.

(b) Routing Layer

This layer determines packet paths through the interconnect. In larger systems, it enables complex routing options based on device population, resource partitioning, and mapping around failing resources due to RAS events.

(c) Transport Layer

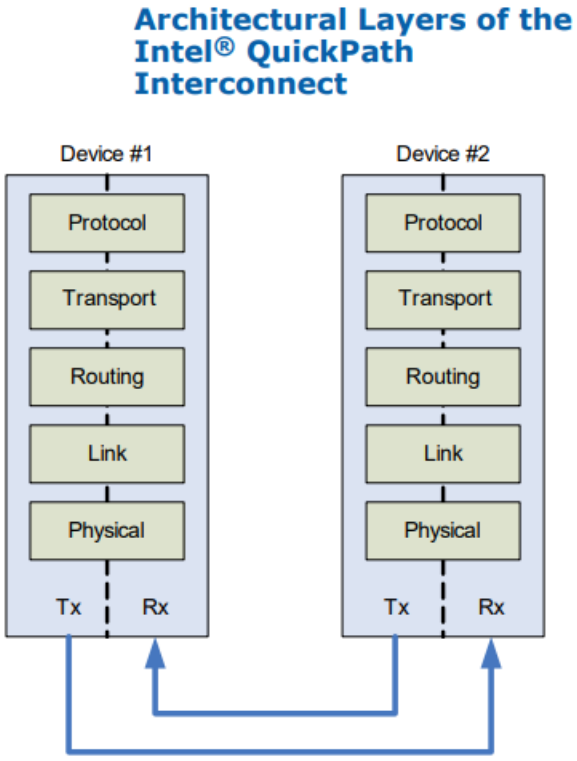
An optional layer ensuring end-to-end transmission reliability. Though not initially a part of Intel's product implementations, it's being considered for future products.

(d) Protocol Layer

Defines the packet transfer units, offering flexibility across different market segment requirements.

(e) Agents

- The Intel QPI includes caching and home agents responsible for managing coherent transactions and memory.
- The interconnect employs two snoop behaviors: home snoop optimized for scalability and source snoop for lower latency in smaller systems.



2. Application

The Intel QPI is predominantly utilized in server-grade applications requiring high-bandwidth, low-latency communication. Its role is crucial in:

- Data Centers: Facilitating communication in large-scale servers with distributed resources.
- Enterprise Computing: Providing high-speed connectivity in mission-critical environments.
- Innovative Silicon Designs: Coupling processors with innovative accelerators or FPGA modules for specialized applications.

3. What Makes It Different

The Intel QPI stands out due to:

- RAS Features: Incorporates error detection, self-healing links, clock fail-over for enhanced reliability.
- Direction Independence: Allows operation in RAS modes without loss of functionality, albeit reduced bandwidth in affected links.
- Snoop Behaviors: Optimized behaviors for scalability (home snoop) and lower latency (source snoop) in different system configurations.

4. Quantifying Performance

Performance which we need:

- Transfer Speed: 6.4 GT/s double-pumped data rate.
- Latency: Transfers a 64-byte cache line in approximately 5.6 ns at 6.4 GT/s.
- Compatibility: Specifically designed for Intel processors, supporting high-bandwidth communication.
- Cost: Varied based on the product and implementation specifics.
- Bandwidth: The interconnect provides a raw bandwidth of 25.6 GB/s and has lower overhead for smaller packet sizes compared to other protocols.
- Protocol Compatibility: Optimized to offer high compatibility and throughput across various market segments, ensuring different application needs are met efficiently.

Processor Interconnect Comparison

	Intel Front-Side Bus[3]	Intel® QuickPath Interconnect[3]
Topology	Bus	Link
Signaling Technology[1]	GTL+	Diff.
Rx Data Sampling[2]	SrcSync	FwdClk
Bus Width (bits)	64	20
Max Data Transfer Width	64	16
Requires Side-band Signals	Yes	No
Total Number of Pins	150	84
Clocks Per Bus	1	1
Bi-directional Bus	Yes	No
Coupling	DC	DC
Requires 8/10-bit encoding	No	No

Contemporary Interconnect Bit Rates

Technology	Bit Rate (GT/s)
Intel front-side bus	1.6[1]
Intel® QuickPath Interconnect	6.4/4.8
Fully-Buffered DIMM	4.0[2]
PCI Express* Gen1[3]	2.5
PCI Express* Gen2[3]	5.0
PCI Express* Gen3[3]	8.0

SUMMARY

With its high-bandwidth, low-latency characteristics, the Intel® QuickPath Interconnect advances the processor bus evolution, unlocking the potential of next-generation microprocessors. The 25.6 GB/s of low-latency bandwidth provides the basic fabric required for distributed shared memory architectures.The inline CRC codes provide more error coverage with less overhead than serial CRC approaches. Features such as lane reversal, polarity reversal, clock fail-over, and self-healing links ease design of highly reliable and available products.The two-hop, source snoop behavior with cache line forwarding offers the shortest request completion in mainstream systems, while the home snoop behavior allows for optimizing highly scalable servers. With all these features and performance it's no surprise that various vendors are designing innovative products around this interconnect technology and that it provides the foundation for the next generation of Intel® microprocessors and many more to come.
For more information refer to the link: https://www.intel.com/content/dam/doc/white-paper/quick-path-interconnect-introduction-paper.pdf

SGI NUMALINK

The SGI NUMAlink network is the system interconnect used within SGI® cache-coherent NUMA (ccNUMA) compute servers. It is highly differentiated from other system interconnects to minimize latency, while providing the extremely high bandwidth and reliability required for High Performance Computing. Paramount among these requirements, is the ability to directly access large cache-coherent memory address spaces with high bandwidth and low latency while maintaining a low bit error rate. NUMAlink was designed with these requirements in mind. From the physical layer where high speed, low latency electrical transfer is accomplished to the routing layer where messages are transferred with low overhead and latency, NUMAlink is optimized for high performance computing.

1. STRUCTURE OF INTERCONNECT

SGI's NUMAlink technology is a high-performance system interconnect designed for SGI's cache-coherent NUMA (ccNUMA) compute servers. It allows for direct access to large cache-coherent memory address spaces with high bandwidth and low latency.

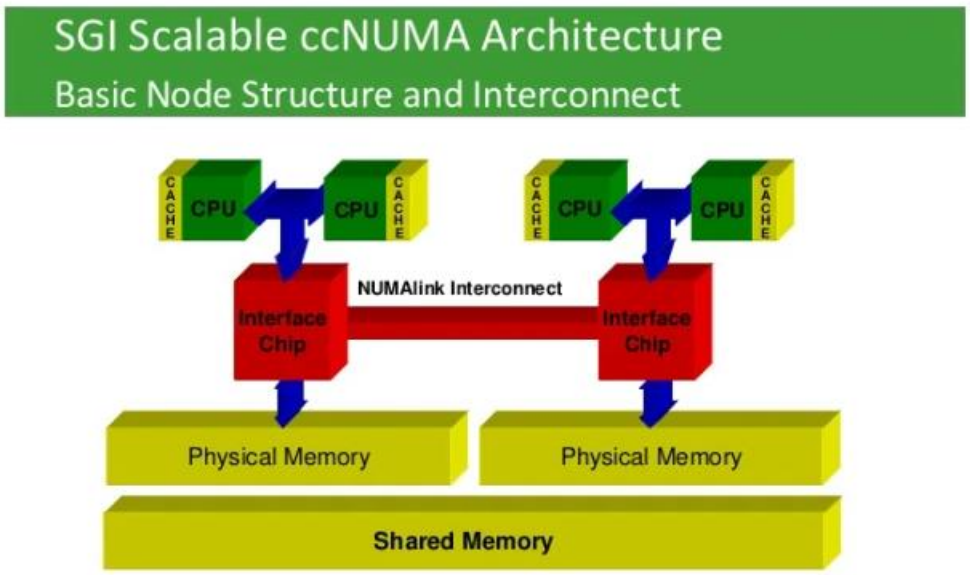
Physical Layer Structure

- NUMAlink 2: Operates at a switching rate of 400Mb/s across 20 data bits/direction. It utilizes a source synchronous clock to capture data, optimized for minimal latency.
- NUMAlink 3: Doubled the switching rate to 800Mb/s, improved on-chip termination resistors, and supported up to 2048 processor sockets with a globally addressable memory space of up to 1TB.
- NUMAlink 4: Uses simultaneous bidirectional signaling at 800Mb/s with each wire pair transporting signals in both directions. It also operates in NUMAlink 3 mode, supporting backward compatibility.

Routing Layer Structure

- NUMAlink's routing layers have been optimized for relatively small message sizes, enhancing the efficiency of cache line transfers.
- Each NUMAlink 4 micropacket contains 16 bytes of data and 4 bytes of link-level sideband, minimizing overhead for small message sizes.

NUMA link 5, currently under development, aims for higher unidirectional signaling, expanded routing layer improvements for multi-paradigm computing, and direct integration of



various computing elements like FPGAs, graphics units, and vector processing units into the NUMAlink memory fabric.

2. APPLICATIONS

Globally Addressable Memory:

NUMAlink facilitates a shared memory, globally addressable system interconnect. SGI's systems offer memory scalability into the petabyte levels, allowing various types of memory to be mapped into a single global address space, unmatched by other interconnects.

Direct Memory Access:

Unlike other interconnects, NUMAlink is directly connected to the memory infrastructure of the system, reducing latency, and enhancing bandwidth by avoiding limitations caused by IO subsystems.

3. What Makes It Different

Unique Advantages of NUMAlink:

- Direct memory access and globally addressable memory reduce latency and enhance bandwidth.
- Eliminates the need for transit through an IO subsystem, offering faster, more direct connections.
- Optimized routing layers specifically designed for efficient cache line transfers.

4. PERFORMANCE QUANTIFICATION

Key Performance Metrics

- Transfer Speed: Ranges from 800Mb/s in NUMAlink 2 to 3.2GB/s in NUMAlink 4, with upcoming NUMAlink 5 promising higher speeds.
- Latency: Ranges from 28ns in NUMAlink 3 and 4 to 55ns in NUMAlink 2. NUMAlink 5 aims to further reduce latency.
- Bandwidth: NUMAlink 4 provides 3.2GB/s/direction.
- Protocol Compatibility: Supports MPI, SHMEM, and direct memory load-and-store approaches, providing efficient memory sharing.
- Cost: Relative cost would depend on factors like development, deployment, and scalability, often reflecting a higher initial investment for greater performance benefits.
- Compatibility: It offers backward compatibility between NUMAlink 4 and NUMAlink 3 modes.

Summary of SGI NUMAlink interconnect generations

Generation	Signaling Rate	Link Bandwidth per Direction	Physical Layer Latency (inc. 3m cable, 24" pcb)	Products/Year of Introduction
NUMAlink 2	400Mb/s	800MB/s	55 ns	Origin 200 and Origin 2000/1997
NUMAlink 3	800Mb/s	1.6GB/s	28 ns	Origin 300 and Origin 3000/2000
NUMAlink 4	1600Mb/s (simul. Bidir)	3.2GB/s	28 ns	Altix/2004

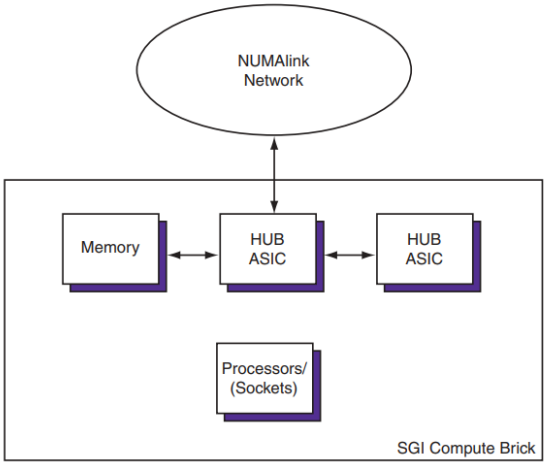
Latency and bandwidth performance of common intereconnect technologies

Technology	Vendor	MPI Latency, µsec, short msg	Bandwidth per Link (Unidirectional, MB/s)
NUMAlink 4 (Altix)	SGI	1	3200
RapidArray (XD1)	Cray	1.8	2000 (1)
QsNet II	Quadrics	2	900 (2)
Infiniband	Voltaire	3.5	830 (3)
High Performance Switch	IBM	5	1000 (4)
Myrinet XP2	Myricom	5.7	495 (5)
SP Switch 2	IBM	18	500 (6)
Ethernet	Various	30	100

Comparison of Linpack system efficiencies in the November 2004 Top 500 list

System/Interconnect	Ave. Linpack Efficiency for 256P System, Percent*	Sample size, number of systems on list*
SGI Altix and NUMAlink	84	14
HP Superdome	79	18
Various/Quadrics	75	4
Various/Infiniband	75	3 (one system at 288P)
Various/Myrinet	63	19
Various/Gigabit Ethernet	59	14

5. INTERLINK CONNECTION



Summary

The SGI NUMAlink network, designed for SGI's cache-coherent NUMA compute servers, stands out due to its focus on minimal latency, high bandwidth, and reliability crucial for High Performance Computing. This system interconnect ensures direct access to extensive cache-coherent memory spaces with low latency and high bandwidth, spanning the Physical Layer (NUMAlink 2, 3, and 4) optimized for speed and low latency, especially in NUMAlink 4, which implements bidirectional signaling for increased performance. Future development with NUMAlink 5 aims for higher signaling rates and expanded routing layers. Notably, NUMAlink's unique advantages lie in its direct memory access, globally addressable memory, elimination of IO subsystem transit, and optimized routing layers, which enable efficient cache line transfers. Performance metrics range from transfer speeds of 800Mb/s to 3.2GB/s, varying latency, and robust protocol compatibility supporting MPI, SHMEM, and direct memory load-and-store approaches, albeit with a potentially higher initial cost for improved performance. The interlink connection within NUMAlink facilitates efficient, low-latency communication between SGI's ccNUMA compute servers, offering a shared memory space and direct connection to the memory infrastructure without IO subsystem limitations.

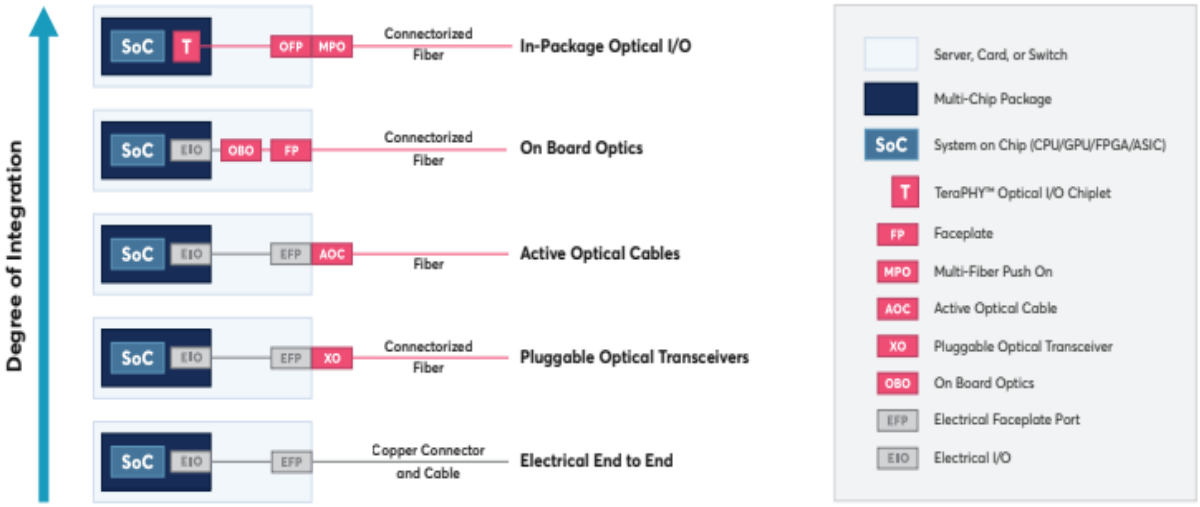
For more information refer to the link :
https://www.cs.ucr.edu/~bhuyan/CS213/2004/numalink.pdf#:~:text=It%20operates%20at%20the%20physical%20layer%20using%20an,signaling%20utilized%20in%20NUMAlink%202%20and%20NUMAlink%203

OPTICAL I/O CHIPLETS

Increasing demand for artificial intelligence, high-performance computing, aerospace data collection and communication, hyper-scale data centers, disaggregated memory and storage has led to the development of a variety of powerful, high-throughput, system-on-chip (SoC) solutions that redefine traditional computing architectures. These new architectures accelerate the data rate of communications between die, sockets, boards, systems, and racks. Electrical I/O has been a bottleneck to scaling processor and SoC performance for the past quarter-century of Moore's Law scaling. It is now a significant barrier to scaling the performance of high-throughput SoC architectures.

Experts agree that electrical SerDes, the most common form of electrical I/O, is hitting a wall. Going beyond 112 gigabits per second (Gbps) is extremely challenging because the large signal losses in copper interconnects at the board level make it hard to transmit data further than a few centimeters at such a high data rate . The next wave of high-performance computing architectures require a new form of universal I/O that eliminate the bottlenecks created by electrical I/O.

The semiconductor industry is moving to chiplets to lower costs, improve yields, enhance reliability, and deliver faster time-to-market for next-generation product designs. Chiplets shorten product cycles by separating the development timelines of different functional building blocks on smaller pieces of silicon instead of putting every function on a single die. This approach allows designers to use advanced process technologies on functionality that requires best available area density and power efficiency, while functionality that does not require the most advanced manufacturing is produced with older, more reliable technology. Chiplet architectures also improve yields by combining smaller pieces of silicon, reducing the risk of manufacturing defects that increase with die size.Industry approach to optical Interconnects



Electrical I/O Barriers to High-Performance Architectures

Electrical I/O can't deliver the power efficiency, latency and bandwidth density required for the next wave of high-performance architectures. For future technology we can't continue with traditional electrical I/O due to following reason :

- Power Efficiency : Power efficiency is a critical limitation when designing electrical systems and datacenters since it directly impacts heat and reliability. Today, the power efficiency of long-reach (LR) electrical I/O at 112 Gbps is 6-to-10 pico Joules per bit (pJ/b). Reaching from the package to the edge of a printed circuit board (PCB) is possible at this data rate but draws a lot of power. Reaching from the package to go across systems, racks and datacenters draws significantly more power, requiring a combination of electrical I/O and pluggable optics.
- Latency : Latency is another critical design factor, limiting the size and number of components interconnected into a system. On-board and off-board electrical I/O at data rates above 50 Gbps require forward error correction (FEC) coding, which introduces added latency of ~100 ns. Such latency, while tolerated in networking applications, is not tolerated in distributed computing systems (memory semantic fabrics) such as those used for machine learning training, inference, and other high-performance computing applications.
- Bandwidth Density : Today electrical I/O provides bandwidth density around 200 Gbps/mm, supporting 25.6-Tbps (terabits per second) Ethernet switch chips. Next-generation 51.2-Tbps Ethernet switch chips will require bandwidth density around 500 Gbps/mm. Future 102.4-Tbps Ethernet switch chips will require bandwidth density around Tbps/mm. While the latest generation of 112-Gbps long-reach electrical SerDes solution can deliver bandwidth density at 200-500 Gbps/mm, there is no roadmap for SerDes technology to achieve bandwidth

densities beyond this limit. Long-reach electrical SerDes at high data rates also suffer from increased package complexity, cooling requirements, and costs.

- Reach : Electrical reach is the distance between bumps inside an MCP or ball-to-ball across the PCB. Table 1 shows reach segmentation by length, loss, and applications. Electrical I/O is suitable for applications within a package but does not scale for connections outside the package due to signal loss, the need for repeaters, and the amount of error correction required. Correcting for these inherent electrical limitations quickly drives power curves beyond sustainable levels.

Length		Loss	Application
< 10 mm/0.4 in	USR	1.5dB@14GHz 3dB@28GHz	Bump-to-bump Inside MCP or 3D Stack
< 50 mm/2.0 in	XSR LPPI	4dB@14GHz 8dB@28GHz	Ball-to-Ball Across PCB
< 200 mm/7.9 in	VSR C2M	10dB@14GHz 20dB@28GHz	Ball-to-Ball
< 500 mm/19.7 in	MR C2C	20dB@14GHz 40dB@28GHz	Ball-to-Ball
< 1000 mm/39.4 in	LR C2F	35dB@14GHz	Ball-to-Ball

MARKET AVAILABILITY

Ayar Labs offering optical interconnect solutions.

Intel also working in this technology. Currently Intel collab with Ayar Labs develop optical interconnect.

CURRENT MARKET STATUS:

Overview of companies offering optical interconnect solutions.

Adoption rates in various industries.

FUTURE PROSPECTS:

Growth potential and expected market trends.

Potential for advancements and improvements in technology.

AYAR LABS IN-PACKAGE OPTICAL I/O CHIPLET

Ayar Labs' solution combines TeraPHY™, an in-package optical I/O chiplet, with SuperNova™, a multiwavelength optical source, to eliminate I/O bottlenecks, transcend process limitations, and unleash innovative architectures. TeraPHY chiplets disrupt the traditional performance, cost, and efficiency curves of the semiconductor and computing industries by combining silicon photonics with standard CMOS manufacturing processes to deliver up to 1000x bandwidth density improvement at 1/10th the power compared to electrical I/O.

Developed in a high-volume GlobalFoundries 45 nanometer process, TeraPHY chiplets integrate millions of transistors with hundreds of photonic devices to drive tens of Tbps of bandwidth up to 2 km out of the package with unmatched power efficiency of less than 5pJ/b. Latency is only 10ns + 5ns/m, point-to-point, with no need for repeaters or FEC, allowing designers to create logically connected, physically distributed compute architectures that scale across racks. TeraPHY delivers bandwidth density more than 200 Gbps/ mm today with a roadmap to Tbps/mm for future generations of high-performance architectures.

Optical I/O Requirement	Ayar Labs TeraPHY™ Performance
Power efficiency	<5pJ/b, roadmap to <2pJ/b
Latency	< 2 x 5ns + TOF
Bandwidth density	In excess of 200 Gbps/mm, roadmap to Tbps/mm
Reach	Package to package connections from mm up to 2 km

Table 2 - Summary of Optical I/O Requirements and TeraPHY Performance

Figure 2 shows an example MCP assembly hosting two TeraPHY chiplets, each delivering up to 2 Tbps of optical I/O bandwidth. Figure 3 shows an opened-lid MCP containing four TeraPHY chiplets integrated into the package with an SoC. MCP technologies (Embedded Interconnect Bridge (EMIB), Silicon-interposer, High-density fanout) provide many advantages over large monolithic dies, including improved yield, support for multiple process nodes in one package, and increased power efficiency.

Figure 2 - Example Multi-Chip Package Assembly with TeraPHY™ Optical I/O

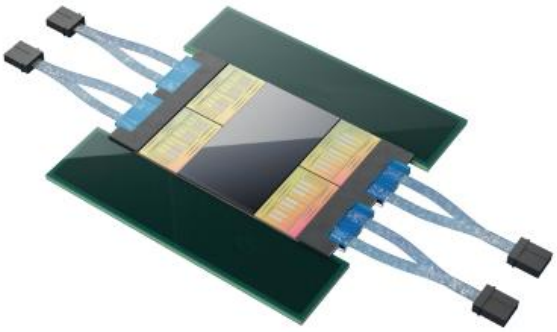


Figure 3 - TeraPHY™ Optical I/O Multi-Chip Package Integration Example

Current pluggable optical solutions undergo full multi-chip packaging (driver chips, silicon photonic chip, laser diodes, etc.) for every optical port (100-400 Gbps). As shown in Figure 4, TeraPHY OIO chiplets (each containing up to eight 256-Gbps optical ports) are flip-chip attached (e.g. C4 or Cu-pillar/µBump) to simplify in-package integration of many optical ports and automate assembly.

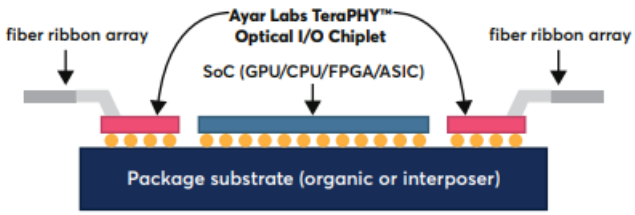


Figure 4 - Flip-Chip Attached for High Volume Manufacturing

Many other configurations are possible, including multiple TeraPHY chiplets in a single package, hosting multiple MCPs on the same board, and hosting SuperNova modules on a separate board or rack to supply light to TeraPHY chiplets on boards distributed across multiple datacenter racks.

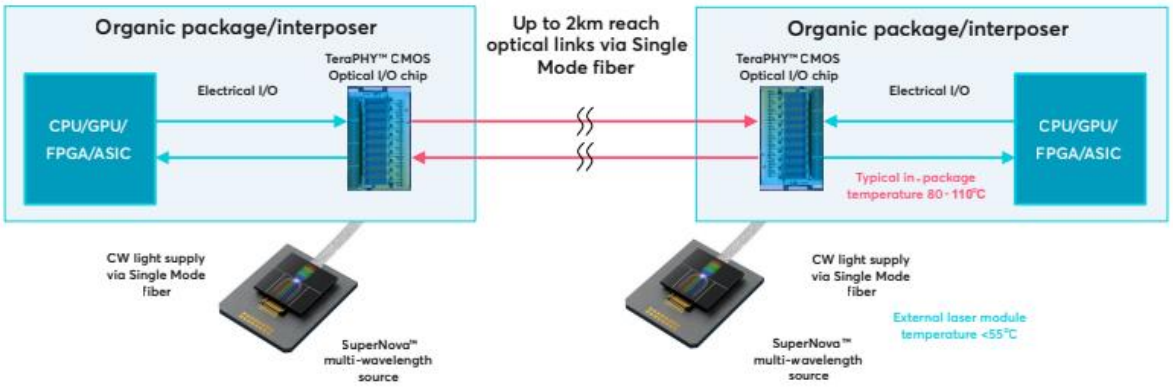


Figure 5 - Example Application of TeraPHY™ Optical I/O

Figure 6 zooms in to show the floorplan for the TeraPHY OIO chiplet. The Advanced Interface Bus (AIB) provides an electrical interconnect to the SoC, operating at an aggregated bandwidth of 2 Tbps . This interface requires 2.5D package integration, which can be realized with a range of technologies such as Embedded

Multi-Die Interconnect Bridge (EMIB) [13], silicon interposer, Chip-on-Wafer-on-Substrate (CoWoS), InFO, and redistribution layer (RDL) [14]. Future TeraPHY chiplets can support other parallel interfaces such as OpenHBI/BoW/proprietary or high-speed serial interfaces such as PCIe/JESD/XSR.

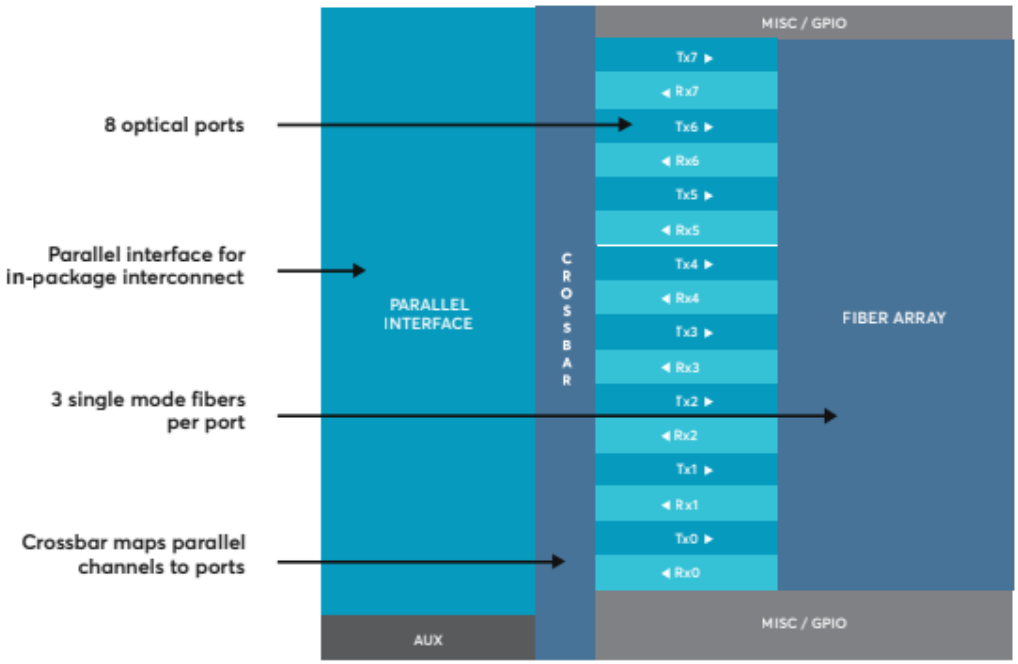


Figure 6 - Ayar Labs TeraPHY™ OIO Chiplet Floorplan

Figure 7 abstracts a single TeraPHY Tx channel and Rx channel, showing microring resonators encoding data from the AIB onto the wavelengths supplied by SuperNova. The encoded light is transmitted up to 2 km to another TeraPHY chiplet, where the corresponding microring resonator on the Rx port converts the light back into an electrical signal. This approach does not require additional protocols, repeaters, or FEC – minimizing both latency and power.

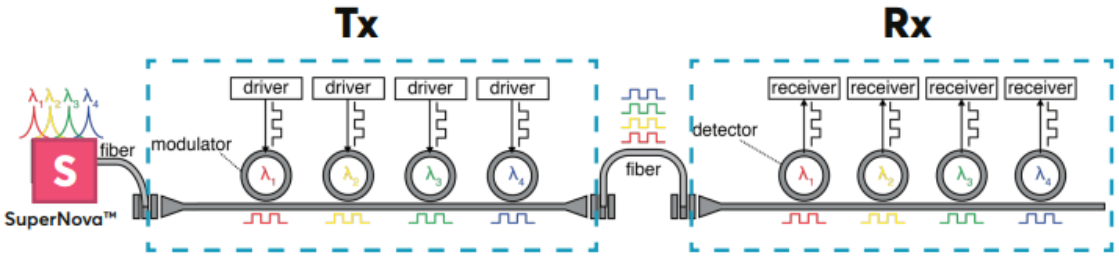


Figure 7 - TeraPHY™ Tx and Rx Channel Abstraction

Port capacity scales with the number of wavelengths per port and data rate per wavelength. Data may be encoded at 16 Gbps/λ, 25 Gbps/λ or 32 Gbps/λ to achieve up to 256 Gbps per optical port

and 2 Tbps per chiplet as shown in Table 3. Ayar Labs has demonstrated technology components providing up to 100 Gbps/λ, paving the way for a long-term bandwidth scaling roadmap.

No. of λ per Optical Port	Gbps per λ	Gbps per Optical Port	Tbps per Chiplet	Bandwidth Density Gbps / mm
8	16	128	1.024	114
8	25	200	1.6	178
8	32	256	2.048	228

Table 3 - TeraPHY™ Data Rates and Bandwidth Density

Ayar Labs has developed SuperNova, the industry's first multi-wavelength, multi-port optical source with 64 wavelengths and the first optical source designed to be compliant with the CW-WDM MSA specification released in 2021. The Ayar Labs SuperNova remote light source (Figure 8) can be deployed across a wide range of applications including high-speed I/O, artificial intelligence, optical computing, and high density, copackaged optics

As the backbone of Ayar Labs' OIO solution, SuperNova provides up to 16 wavelengths of light, powering up to 16 ports, and is capable of supplying light for 256 channels of data (or 8.192 Tbps at full capacity). The solution provides up to 1000x the bandwidth at 1/10th of the power compared to electrical I/O alternatives – effectively eliminating I/O bottlenecks and transcending process limitations.

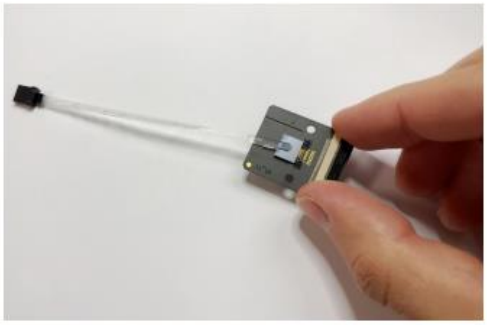


Figure 8 – Ayar Labs SuperNova™ Optical Source

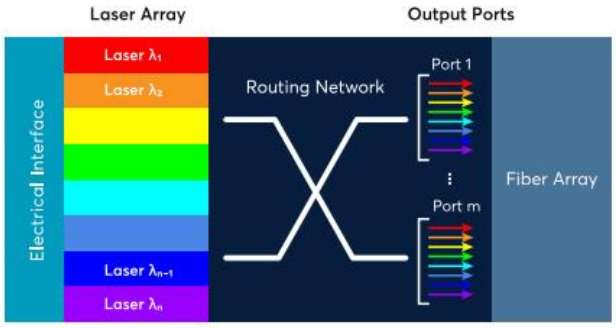


Figure 9 - SuperNova™ Optical Source Floorplan

ROAD TO COMMERCIALIZATION FOR OPTICAL CHIP-TO-CHIP INTERCONNECTS

The communications industry has an I/O bottleneck problem. When data is unable to be moved into or out of a chip or chiplet at a rate that matches or exceeds the rate at which processing occurs, a backlog is created in the data stream, such that processing/storage conducted in other chips or chiplets is stalled while they await the data. The I/O bottleneck has

implications for power consumption in that a significant portion of the power that should be spent on computation and processing is instead used to push data through the system.

MORE DATA, MORE QUICKLY

The I/O bottleneck problem is rooted in the electrical interconnects used between chips. The conductive traces—by virtue of passing electrical currents—experience resistance, which depends on such factors as the length and cross-sectional area of the trace. The resistance has a direct impact on the speed at which the signal is transmitted between chips. While resistance in electrical traces has previously been acknowledged, for the most part, it hasn't needed to be addressed—until now.

A disadvantage of advancing technologies is that, sooner or later, the issue becomes unavoidable if further progress is to be made. Such is the case for electrical interconnects, in which chip and package densities have become small enough—and applications demanding enough—that electrical interconnects are reaching the end of their scalability; the detriment to performance is becoming unavoidable.

Optical interconnects are presented as a solution to these performance issues. Using light as the data signal between chips allows for higher-bandwidth transmissions than for electrical interconnects because photons (the quanta of light) are massless and thus do not experience resistance in the same way that electrons do. While photons can be lost via material absorption, scattering with the waveguide and other effects, those occurrences can be mitigated through a careful choice of materials, enabling optical interconnects' data-transmission rates to exceed those possible with electrical interconnects.

ADOPTION CHALLENGES & OPPORTUNITIES

Chip manufacturing—encompassing all the steps necessary to imprint integrated-circuit designs on semiconductor wafer—is an established process that has been optimized to minimize materials losses, manufacturing time and equipment degradation. The use of chip-to-chip optical interconnects requires a change to this process because it requires conversion of the electrical signal to an optical signal at the chip-edge I/Os. This would typically be achieved by incorporating lasers or LEDs into the process—an action that requires consideration of the current required for correct functioning (the lasing threshold for lasers), in addition to whether any other optoelectrical components are required. The inclusion of optoelectrical components into a nominally electrical workflow is a significant change that is reflected in the incurred developmental costs.

Fortunately, developments in several areas promise to lower the barrier to optical interconnect adoption by reducing costs. The relaxing of alignment tolerances is one key way of

lowering costs at the assembly stage. Companies working on technology to enable this include Avicena, which is using multimode fibers rather than single-mode fibers (the tradeoff is that you would have to use LEDs rather than lasers), and Teramount, which uses micro-optical elements it calls PhotonicBumps to allow for vertical coupling (rather than coupling at the edge of the die).

In another vein, an increased throughput-to-power-consumption ratio presents savings for the end user. Here, a move to dense-wavelength–division multiplexing (DWDM) is seen as instrumental to achieving higher rates of data transmission, as it allows more wavelengths to be transmitted over a single channel. Intel used DWDM via a distributed-feedback laser in a silicon photonics co-packaged pluggable optics module in June 2022. And Quintessent is looking to achieve coarse WDM (CWDM) × DWDM by using quantum dots to realize a comb laser for 32 wavelengths in a single fiber (what would otherwise require multiple tunable distributed-feedback lasers to achieve).

With the progress made in high-performance computing, AI and 5G communications over the past couple of years, a handful of startups specializing in addressing the I/O bottleneck via optical implementations have received significant funding from major players. Chief among these is Ayar Labs, which has developed an optical I/O chiplet built in a 45-nm process by GlobalFoundries. The product, TeraPHY, can reach a bandwidth density of 200 Gbps/mm at the time of writing, with a roadmap to achieving 1 Tbps/mm of bandwidth, which would push optical interconnects firmly into the realm of

industry-leading bandwidth density.

Major companies like Intel, HP and Nvidia are watching closely as Ayar Labs' technology is put to the test across various settings, from creating an AI infrastructure using optical I/O to co-packaged pluggable technology. Both Intel and Hewlett Packard Enterprise have invested in Ayar Labs.

Reference : EE Times Europe https://www.eetimes.eu/road-to-commercialization-for-optical-chip-to-chip-interconnects/

THE EARLY STAGES OF COMMERCIALIZATION

Whether optical interconnects can be truly commercialized remains an open question, given the changes this would bring to the mature semiconductor industry and the need for collaboration within the supply chain.

The view of IDTechEx is that optical I/O will be commercialized but that the use cases will be largely restricted to data center use. This is because edge devices are not yet dealing with the types of compute-intensive workloads that would benefit from architectural disaggregation (in which high-bandwidth memory sits off-package and is connected to compute clusters via optical interconnects so that compute can be effectively deployed and power efficiency optimized). Notable exceptions would be autonomous vehicles and military sensing applications, in which low latency is critical. The benefits of optical I/O become harder to ignore the further down the line one gets for transmission rates and AI models. To ensure cost-effectiveness at higher transmission rates and more complex models, optical I/O is an attractive solution.

CONCLUSION

Semiconductor designers are creating new, chiplet based architectures that deliver the second phase of Moore's Law to lower costs, improve yields, enhance reliability, and deliver faster time-to-market for nextgeneration designs. At the same time, the limitations of electrical I/O are driving the search for a new form of universal I/O that quickly delivers terabit data rates from die-to-die and across data centers with less power. TeraPHY OIO chiplets eliminate electrical I/O bottlenecks and transcend process limitations to unleash the next wave of innovation in semiconductor and datacenter design.

Ayar Labs is the first to deliver monolithic in-package optical I/O chiplets, a new universal I/O solution that replaces traditional electrical I/O and enables chips to communicate with each other from millimeters to kilometers, to deliver orders of magnitude improvements in latency, bandwidth density, and power consumption.

References

1. A. Steffora Mutschler, "Wrestling With High-Speed SerDes," Semiconductor Engineering, 2019. [Online]. Available: https://semiengineering.com/wrestling-with-high-speed-serdes/ .
2. G. E. Moore, "Cramming more components onto integrated circuits. In: Electronics," Electronics, 1965.
3. L. T. Su, S. Naffziger, and M. Papermaster, "Multi-chip technologies to unleash computing performance gains over the next decade," in Technical Digest - International Electron Devices Meeting, IEDM, 2018.
4. T. Simonite, "To Keep Pace With Moore's Law, Chipmakers Turn to 'Chiplets,'" Wired, 2018. [Online]. Available: https://www.wired.com/story/keep-pace-moores-law-chipmakers-turn-chiplets/ .
5. T. Coughlin, "Chiplets for All," Forbes, 2019. [Online]. Available: https://www.forbes.com/sites/tomcoughlin/2019/05/11/chiplets-for-all .
6. E. Sperling, "Chiplets, Faster Interconnects, More Efficiency," Semiconductor Engineering, 2019. [Online]. Available: https://semiengineering.com/chiplets-faster-interconnects-and-more-efficiency/ .
7. C. Sun et al., "Single-chip microprocessor that communicates directly using light," Nature, 2015.
8. V. Stojanović et al., "Monolithic silicon-photonic platforms in state-of-the-art CMOS SOI processes [Invited]," Opt. Express, vol. 26, no. 10, p. 13106, 2018.

9. M. Wade, M. Davenport, M. Rust, and F. Sedgwick, "TeraPHY : A Chiplet Technology for Low-Power , High-Bandwidth In-Package Optical I / O," Hot Chips Conference, 2019. [Online]. Available: https://ayarlabs.com/teraphy-a-chiplet-technology-for-low-power-high-bandwidth-in-package-optical-i-o/ .
10. R. Danilak, "Why Energy Is A Big And Rapidly Growing Problem For Data Centers," Forbes, 2017. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2017/12/15/why-energy-is-a-big-and-rapidly-growing-problem-for-data-centers .
11. J. Goergen, C. Systems, and V. Parthasarathy, "Spanning SERDES Across Reaches - Finding the Best Modulation Approach," no. November, 2014.
12. R. Meade et al., "TeraPHY: A High-Density Electronic-Photonic Chiplet for Optical I/O from a Multi-Chip Module," 2019 Opt. Fiber Commun. Conf. Exhib. OFC 2019 - Proc., no. Mcm, pp. 5–7, 2019.
13. D. Kehlet, "Accelerating Innovation Through A Standard Chiplet Interface : The Advanced Interface Bus (AIB)," 2019
14. R. Mahajan et al., "Embedded Multi-die Interconnect Bridge (EMIB)-A High Density, High Bandwidth Packaging Interconnect," in Proceedings - Electronic Components and Technology Conference, 2016.
15. M. Lapedus, "Bridges Vs. Interposers," Semiconductor Engineering, 2018. [Online]. Available: https://semiengineering.com/using-silicon-bridges-in-packages/ .

Executive Summary

Conventional digital technology is unable to handle the high compute requirements of AI models that need to run in real-time on cost-effective, lowpower hardware. This is a problem that needs to be addressed by a change in technology--a change to a hardware platform that employs analog computein-memory (CIM). This is the technology pioneered by Mythic, a new generation of AI technology that can propel the next wave of applications that harness deep learning and its many variants.

The challenge of inferencing

Conventional digital processing pipelines can support the throughput required of real-time AI, but the platforms come with high costs in terms of silicon real estate and energy. The energy density of the server blades and compute accelerators is reaching levels where operators have been pushed to employ increasingly complex and costly heat-containment and management strategies. Some have begun to investigate the use of liquid cooling, even to the point of immersing the systems in fluids with high thermal conductivity.

For many mobile and real-time industrial systems, pushing inferencing to the cloud is impractical. The systems may not have access to sufficiently highbandwidth communications links, and even where they are available, the latency incurred by passing data over long distances is a major challenge to responsiveness. Real-time systems cannot afford to have image-recognition functions miss deadlines. They need to have these services provided locally.

Key to the high-volume deployment of inferencing systems at the edge is the ability to reduce the energy demand, as well as the amount of silicon area needed. The metric that matters most is "inferences per watt per dollar" (I/W/$): how many image frames can be processed in each amount of time and at what energy and capital cost.

Digital solutions hit the wall

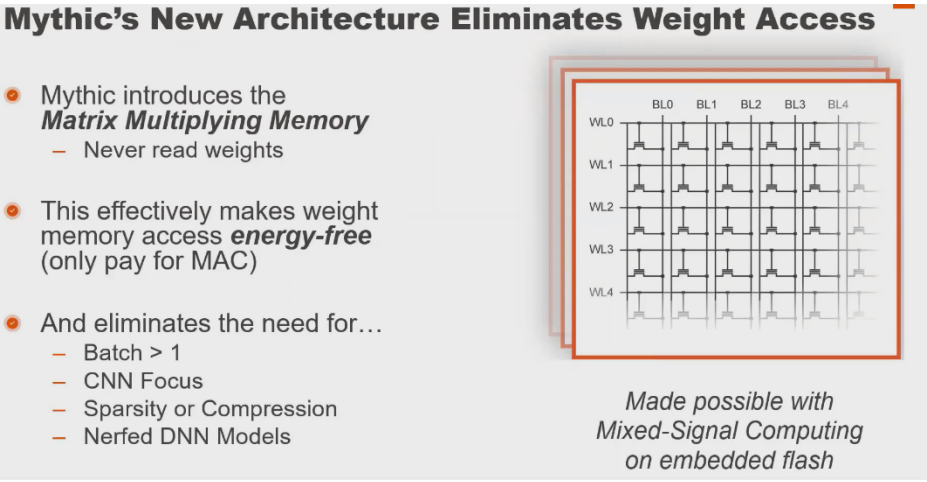
Digital architectures like GPUs are optimized for matrix operations, particularly suited for deep learning. However, their inefficiency in energy consumption arises from two main factors. Firstly, the exponential growth of logic gates required for multiplication in high-throughput operations. To mitigate this, reducing precision, like converting weights to 8-bit values during inference, has been adopted, but further approximations, such as binary or ternary calculations, can compromise model accuracy.

The second challenge involves the inability of digital architectures to effectively utilize the inherent spatial and temporal locality of information in deep learning calculations. Caching input data close to the processor helps reuse these elements multiple times, while the vast number of individual weights required incurs a massive energy burden due to frequent access to remote memory.

Pruning, a post-training process removing less impactful neuron connections, is one technique to reduce weight accesses. However, using both approximation and pruning significantly complicates neural network development, leading to decreased result quality and increased errors. Models must be extensively re-analyzed as pruned versions deviate from those trained at full resolution, adding to development complexity.

An analog compute-in-memory solution

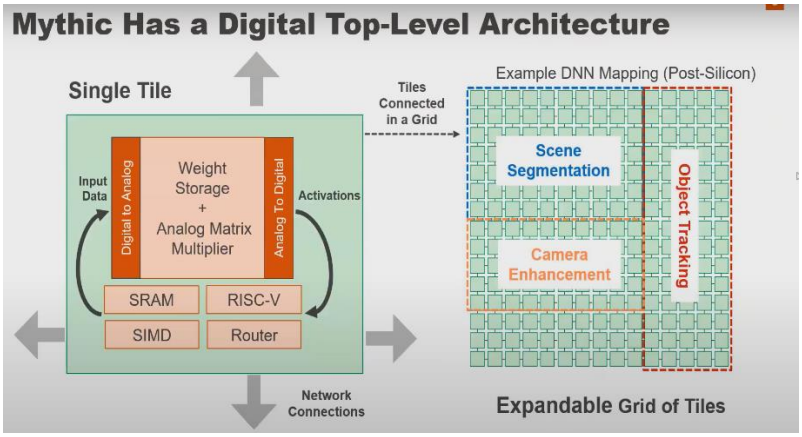
What is required for inferencing at the edge and for low-energy data-center inferencing is an architecture that allows for the seamless transfer of cloudtrained models, but which does not suffer from the high energy cost of memory transfers and high-resolution computation that is common to today's digital implementations. Mythic's compute-in-memory solution uses the memory itself to perform computations and so avoids the need to continually move weight data around the system. This slashes the energy per MAC from 10pJ in a typical digital edge inferencing implementation that holds the large weight arrays in DRAM to as low as 0.5pJ. Across the billions of MAC computations required for video-rate inferencing at moderate resolution, the resulting energy savings are dramatic.



In Mythic's architecture, each access to weight memory is essentially energy-free. The main contribution to the processing energy comes from the MAC operation itself, which is implemented simply by passing data through the memory. As a result, this process is much more efficient than that found in logic-hungry digital datapaths. The core of the Mythic

implementation is a flash memory technology that is widely used in microcontrollers and other mass-production embedded systems.

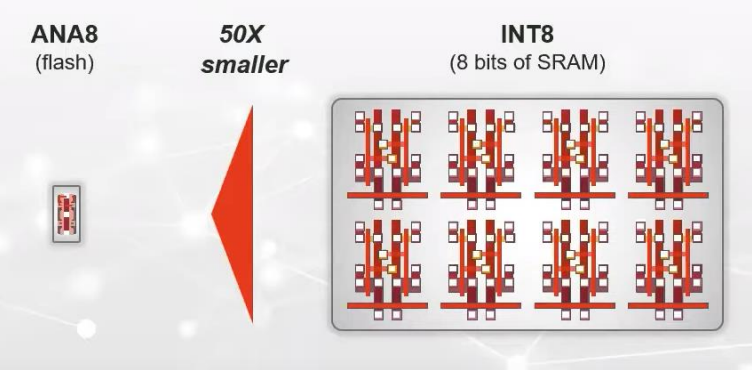
The readout circuitry in a conventional flash memory discretizes these programmed charge values into a 1, 2 or 3-bit value. In practice it is possible, in a flash memory made on a mature process such as 40nm, to store reliably a range of charges that correspond to a digital resolution of 8 bits. A further advantage of the flash memory cell is that it is non-volatile. Once programmed it can store an electrical charge for long periods of time without power, a key advantage for systems that need to run off a battery. But the memory can be erased and reprogrammed at any time to support new or updated models.



When a charge is programmed into a flash memory device, its electric field influences any signal passing through it. In the Mythic architecture, the flash transistor acts as a variable resistor that reduces the signal level passing to the output. That reduction is proportional to the analog value stored in the memory. This simple effect implements the multiplication stage found in DNN calculations. The accumulation process, in which the output from each of those calculations is summed, is handled by aggregating the output of an entire column of memory cells. Thanks to these two properties, the Mythic architecture can process an entire input vector in a single step rather than iterating at high speed as in a digital processor.

Thanks to its ability to streamline MAC processing by orders of magnitude compared to conventional digital processors, Mythic's memory-based accelerator makes it possible to transfer fully trained neural networks directly to a low-energy inferencing platform without any need for pruning or further approximation. But the analog memory-based core is only part of the solution. Functions such as activations and pooling, which form key parts of any DNN are generally best implemented in digital logic. The Mythic architecture accommodates this by including a single-instruction, multiple-data (SIMD) accelerator unit coupled to a RISC-V processor that coordinates operations and local SRAM to hold temporary data. With these components, the Mythic solution can run a complete DNN model independently.

A key element of the Mythic solution is scalability. The combination of memory-based accelerator, SIMD engine, SRAM and RISC-V processor forms a tile in an array of DNN engines. All the tiles are linked by a high-speed network-on-chip (NoC) routing mesh to allow for the efficient flow of input, output, and intermediate data elements. The array of tiles is managed by an on-chip processor and communicates with the system's host processor over a PCIe interface.



The mesh architecture of the Mythic platform provides the ideal substrate for applications such as machine learning. In contrast to the many applications written for conventional architectures, which revolve around the sequential processing of a single stream of code, AI inferencing is a graphbased application. Graph applications are well suited to dataflow architectures where it is straightforward to assign a different compute element to each node of the graph. When one graph node has completed its work, the 6 data output flows to the next graph node for processing. With its combination of different types of compute functions in a mesh, Mythic implements a highly efficient dataflow architecture.

The dataflow architecture also maximizes inference performance by having many of the compute-in-memory elements operating in parallel, pipelining the image processing by handling neural-network layers in parallel in different tiles of the array. By being built from the ground-up as a dataflow architecture, the Mythic solution minimizes the memory and computational overhead required to manage the dependency graphs that define dataflow computing, and it keeps the application operating at maximum performance. The result is an architecture that delivers inferencing performance at breakthrough silicon cost and power levels, while also supporting a wide range of edge and data-center systems.

A new wave of applications

Mythic's energy-efficient architecture enables a broader range of smart devices independent of cloud reliance, beneficial for applications like security cameras, robots, and drones operating for extended periods without fixed power sources. This technology opens doors for manufacturers, particularly in industrial control systems where traditional compute platforms struggle due to high costs and power consumption. Deep learning, facilitated by Mythic's platform, empowers compact, low-power self-contained systems for control processes, offering early anomaly detection in systems with long time constants.

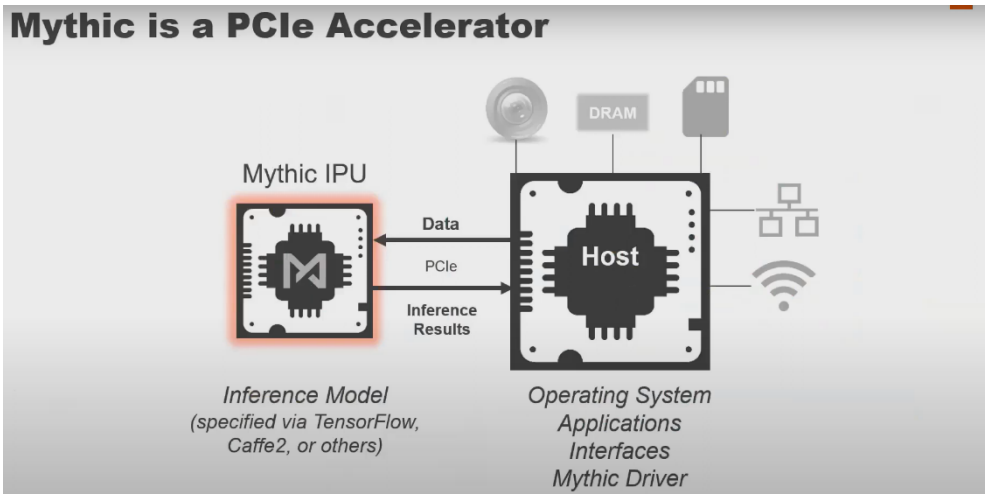
In robotic systems, such as drones, the platform enhances situational awareness by enabling deep learning for image and sensor processing. This capability allows drones to autonomously classify objects, reducing the need for constant human control and enabling applications like infrastructure inspection over long distances. Similarly, within security cameras, Mythic's technology enables on-device anomaly detection, significantly reducing network bandwidth and power requirements, even supporting battery-powered wireless cameras for remote placements.

The architecture's non-volatile flash memory preserves neural network weights, enabling devices to power down when deep learning functions aren't needed, saving energy, and maintaining high battery autonomy. The scalable nature of Mythic's platform allows tailored implementation for specific models, dedicating different tile groups to various tasks within devices like drones or robots. Furthermore, in the data-center environment, Mythic's interchip communication facilitates large-scale inferencing engines with cost and energy efficiency advantages over existing platforms. Its core MAC operations consume significantly less power than traditional platforms, allowing high processing densities without advanced cooling or power infrastructure. The platform scales seamlessly from edge sensor nodes to sophisticated cloud-based DNN inferencing.

Delivering practical inferencing

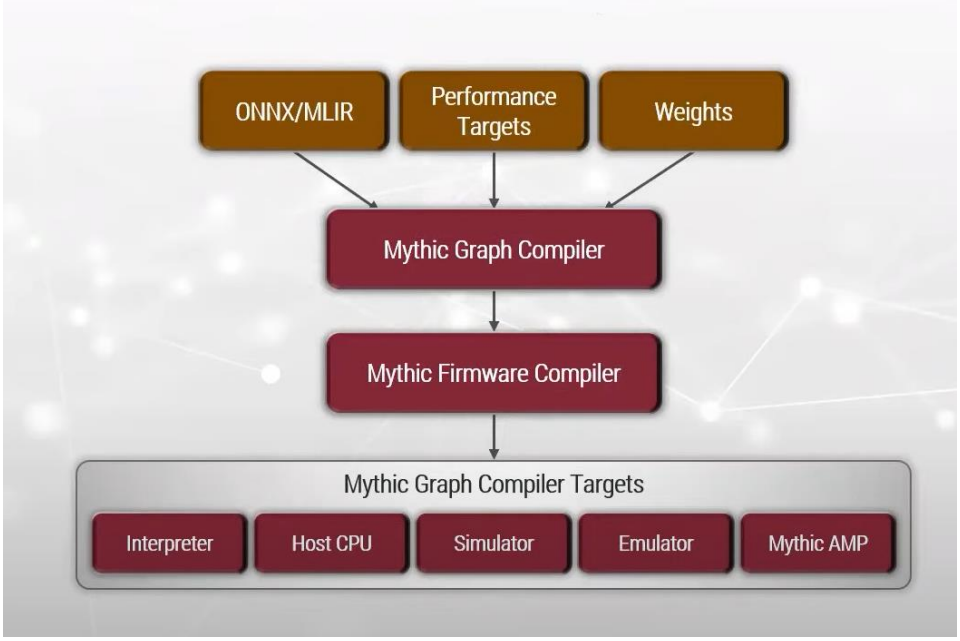
Mythic's platform enables the rapid development of deep-learning applications, as it supports models trained using readily available frameworks, such as ONNX and TensorFlow, without requiring that the developer perform complex runtime

optimizations. Mythic has adopted a software strategy that will ensure the process of transferring models to the end device is as seamless as possible using tools that convert the trained models produced by a variety of machine-learning frameworks into executables that can be downloaded to a single Mythic product or an array of them. Integration into end applications will be supported by libraries that provide application programming interfaces (APIs) to the software running 8 inside the device, making it easy to employ the Mythic silicon as a coprocessor to conventional microcontroller-based designs.



The software developed for the Mythic platform optimizes the neural network for execution through a simple two-stage process. The first step, using the Mythic Optimization Suite, transforms the trained network into a form that is compatible with analog CIM, including quantization from floating-point representations to 8-bit integer. This suite checks that the quantization will not degrade performance below acceptable levels. The tools also include retraining flows for applications that have strict accuracy requirements or more aggressive performance and power targets or a combination of all three. Quantization-aware and analog-aware retraining builds resiliency into layers that are more sensitive to the lower bit-depths of quantization and to analog noise.

The Mythic Graph Compiler follows and performs automated mapping onto the target array, packing, and support-code generation. The output is a packaged binary that contains everything that the host driver needs to program and control the Mythic device to perform inferencing in a real-time environment. Longer term, Mythic envisions an SDK with a suite of powerful tools to help developers evaluate tradeoffs and identify the best solution within the constraints of power and cost.



As well as integrating easily with today's popular tools, it is important to be able to take advantage of the rapid pace of development in the field of machine learning. As new layer types and network topologies are invented, software and hardware support should be straightforward. The Mythic platform ensures this through the modular design of the software SDK, leveraging a large amount of generic matrix compute capabilities rather than architecture-specific accelerators.

Flexibility is just as important to the hardware architecture. The scalable, tiled nature of the Mythic platform greatly eases development and integration into an end system. During prototyping, it is likely that there will be multiple iterations of model tuning and training to ensure that the core deep-learning engine can deal with real-world problems and is unaffected by problems that may be caused by a skewed training set. For example, if the prototyping stage determines that additional models are needed to handle changes in lighting or environmental conditions, a move to a larger array can easily be made. Similarly, the developers may find opportunities for cost reduction that allow the use of a smaller model and a more appropriate device.

Mythic intends to provide a range of implementations that scale in size and number. The platform will be made available not just in IC form but as accelerator cards that may use multiple devices. The tiled architecture eases the production of derivatives based on market demand and so supports the evolution of deep learning in edge and low-power data-center systems as new applications emerge.

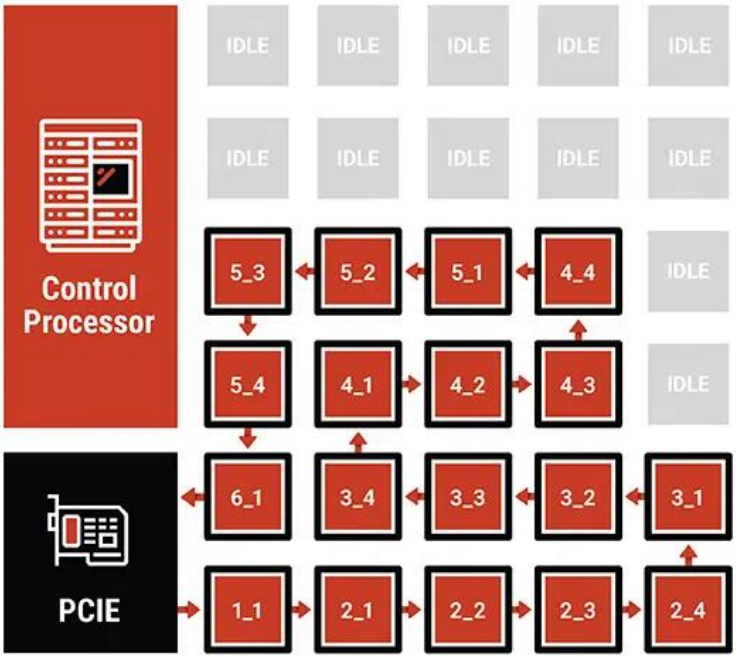
Dataflow Architecture

Standard compute architectures are designed to tackle sequential algorithms. They excel at these algorithms by using a massively powerful and power-intensive CPU core, surrounding it with a memory architecture that matches the memory profile of those applications. AI inference is not a typical sequential application; it is a graph-based application where the output of one graph node flows to the input of other graph nodes.

Graph applications provide opportunities to extract parallelism by assigning a different compute element to each node of the graph. When the results from one graph node are completed, they flow to the next graph node to start the next operation, which is ideal for dataflow architecture. In our dataflow architecture, we assign a graph node to each compute-in-memory array and put the weight data for that graph node into that memory array. When the input data for that graph node is ready, it flows to the correct location, adjacent to the memory array, and then is executed upon by the local compute and

memory. Many inference applications use operations like convolution, which processes bits of the image frame instead of the whole frame at once.

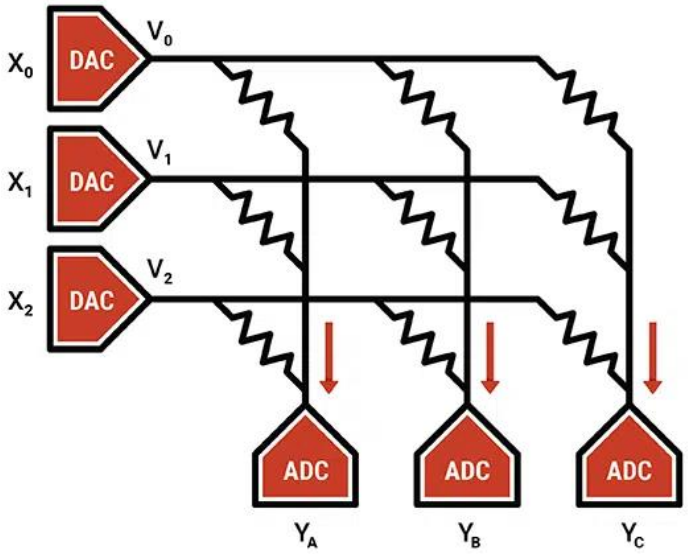
Our dataflow architecture also maximizes inference performance by having many of the compute-in-memory elements operating in parallel, pipelining the image processing by processing neural networks nodes (or "layers") in parallel in different parts of the frame. By being built from the ground up as a dataflow architecture, the Mythic architecture minimizes the memory and computational overhead required to manage the dependency graphs needed for dataflow computing, and keeps the application operating at maximum performance.



Analog Computing

Analog computing provides the ultimate compute-in-memory processing element. The term compute-in-memory is used very broadly and can mean many things. Our analog compute takes compute-in-memory to an extreme, where we compute directly inside the memory array itself. This is possible by using the memory elements as tunable resistors, supplying the inputs as voltages, and collecting the outputs as currents. We use analog computing for our core neural network matrix operations, where we are multiplying an input vector by a weight matrix.

Analog computing provides several key advantages. First, it is amazingly efficient; it eliminates memory movement for the neural network weights since they are used in place as resistors. Second, it is high performance; there are hundreds of thousands of multiply-accumulate operations occurring in parallel when we perform one of these vector operations. Given these two properties, analog computing is the core of our high-performance yet highly-efficient system.



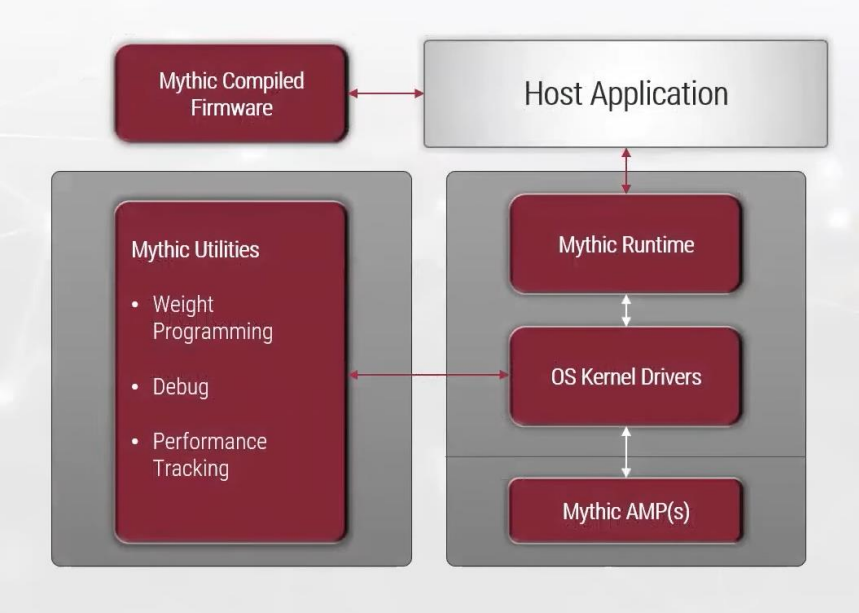
Mythic AI Workflow

Mythic's platform delivers category-leading performance, power, and on-chip model capacity in a cost-effective form factor. To leverage this compute, Mythic's software stack optimizes and compiles trained neural networks using a flow that is familiar and easy to use for developers. We build on existing ecosystems, like PyTorch, for the front-end to ensure frictionless integration with standard training flows.

The software then runs through two stages: optimization and compilation. The Mythic Optimization Suite transforms the neural network into a form that is compatible with analog compute-in-memory, including quantization from floating point values to integer 8-bit. The Mythic Graph Compiler performs automatic mapping, packing, and code generation. The result is a packaged binary containing everything that the host driver needs to program the Mythic AMP™ and run neural networks in a real-time environment.

MYTHIC OPTIMIZATION SUITE

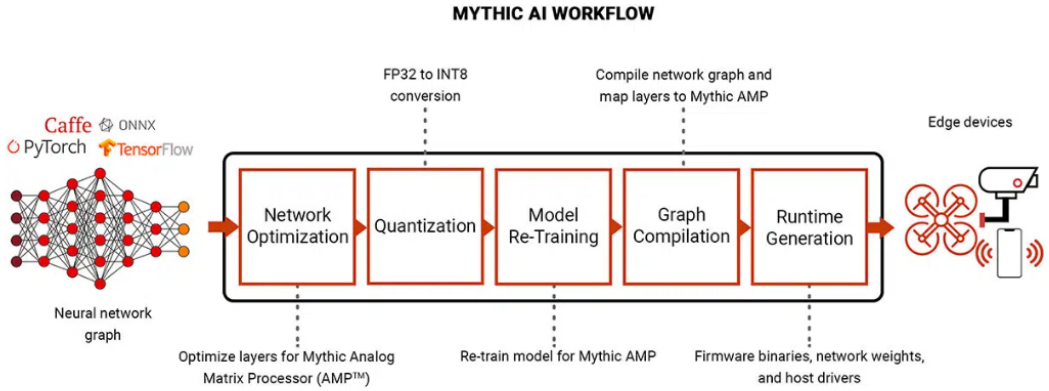
The first stage of the Mythic AI workflow is optimization of a trained neural network. The quantization flow converts 32-bit floating point weights and activations — the standard numerical format in training — to 8-bit integer, which is essential for effective deployment at the edge. Quantization represents a major pain point for customers with high accuracy requirements. Our simple flow runs after training and performs conversion to an analog representation of 8-bit (ANA8). The resulting accuracy is comparable to digital 8-bit quantization which is typically deployed in power-constrained edge applications.



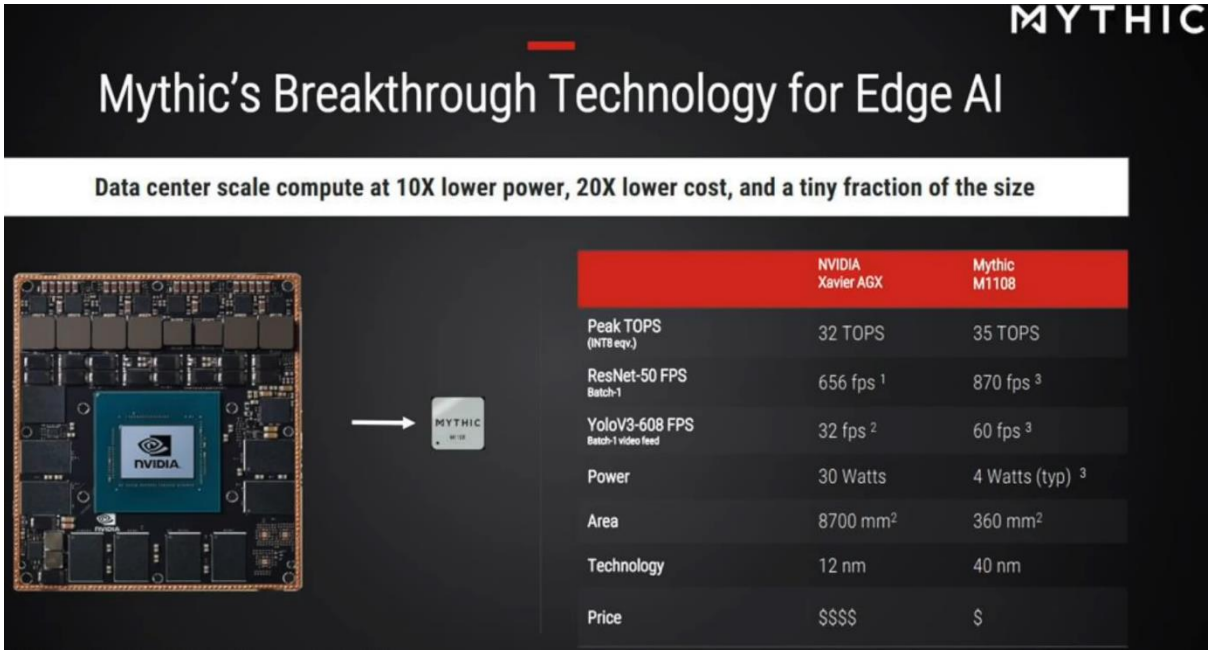
We also provide retraining flows for applications with strict accuracy requirements and/or more aggressive performance and power targets. Quantization-aware and analog-aware retraining builds resiliency into layers that are more sensitive to the lower bit-depths of quantization and to analog effects. For aggressive performance and power targets, certain layers can even be pushed to 4 bits and below without a significant drop in accuracy.

MYTHIC GRAPH COMPILER

The second stage of the Mythic AI workflow generates the binary image to run on our Mythic AMP. The conversion from a neural network compute graph to machine code is handled in an automated series of steps including mapping, optimization, and code generation. Powerful hardware architecture elements including many-core processing, SIMD vector engines, and dataflow schedulers are all leveraged automatically by the graph compiler. Even the host driver is simple and pain-free, with input/output and memory transfers handled behind the scenes. We also provide compiler support for high compute intensity operations before, after, and in-between neural network layers with our array of processors and SIMD vector engines.



Comparison



Conclusion

Mythic's groundbreaking analog compute-in-memory (CIM) architecture revolutionizes the landscape of AI inferencing, offering a transformative solution to the limitations of conventional digital processing in high-compute, real-time scenarios. By leveraging the inherent efficiencies of analog computing within its flash memory-based core, Mythic tackles the fundamental challenges of energy consumption and silicon area, providing a solution that operates seamlessly at the edge while significantly reducing energy demands and capital costs. The platform's scalability, dataflow architecture, and integration ease, combined with a software stack optimized for rapid deployment, not only deliver breakthrough performance and power efficiency but also unlock a new wave of applications across diverse sectors, from robotics and security systems to industrial control and beyond. Mythic's technology marks a pivotal shift in the field, enabling the deployment of deep learning applications at an unprecedented scale, empowering devices to operate independently and efficiently without reliance on the cloud.

BIBLIOGRAPHY

- https://www.nytimes.com/2023/05/11/technology/us-chiplets-tech.html
- Automotive software electronics market 2030 | McKinsey
- https://blackberry.qnx.com/en/ultimate-guides/software-defined-vehicle
- https://www.forbesindia.com/blog/technology/driving-the-future-software-defined-vehicles-and-the-dawn-of-digital-mobility/

Task 1.1

- Designing C-V2X Communication Systems: Key Engineering Considerations and Best Practices (wevolver.com)
- Introduction to Cellular V2X (qualcomm.com)
- Timeline of V2X evolution. | Download Scientific Diagram (researchgate.net)
- https://www.emqx.com/en/blog/what-is-v2x-and-the-future-of-vehicle-to-everything-connectivity
- https://www.eetimes.eu/v2x-communication-benefits-and-limits/
- https://www.jdpower.com/cars/shopping-guides/levels-of-autonomous-driving-explained
  https://www.businesstoday.in/technology/news/story/mercedes-beats-tesla-in-self-driving-becomes-first-certified-level-3-autonomous-car-company-in-us-367937-2023-01-28
- Human-Machine Interaction for Autonomous Vehicles: A Review | SpringerLink
- MediaTek partners with NVIDIA to provide product roadmap for connected cars, ET Auto (indiatimes.com)

Intel:

- https://ieeexplore.ieee.org/document/8726257
- https://ieeexplore.ieee.org/document/7545486
- https://www.anandtech.com/show/16823/intel-accelerated-offensive-process-roadmap-updates-to-10nm-7nm-4nm-3nm-20a-18a-packaging-foundry-emib-foveros/4
- https://ieeexplore.ieee.org/document/8993637
- https://fuse.wikichip.org/news/5949/intel-unveils-foveros-omni-and-foveros-direct-leveraging-hybrid-bonding/
- Cramming More Components onto Integrated Circuits -GORDON E. MOORE

TSMC

- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7545402
- https://ieeexplore.ieee.org/document/8811194
- https://3dfabric.tsmc.com/english/dedicatedFoundry/technology/SoIC.htm

2.5D and 3D Interconnects

- https://www.einfochips.com/blog/2-5d-3d-ics-new-paradigms-in-asic/
- https://www.esda.org/news/what-exactly-is-esd-for-3d-ics

OCP & AXI

- https://www.accellera.org/downloads/standards/ocp
- https://www.xilinx.com/products/intellectual-property/axi_interconnect.html
- https://fpgaemu.readthedocs.io/en/latest/axi.html
- https://developer.arm.com/documentation/102202/0300/AXI-protocol-overview
- Sources: https://www.imec-int.com/en/articles/why-are-chiplets-attracting-attention-automotive-industry
- https://research.nvidia.com/sites/default/files/publications/ISCA_2017_MCMGPU.pdf
- https://www.lantekcorp.com/how-chiplets-may-help-the-future-of-semiconductor-technology/#:~:text=You%20can%20increase%20the%20yield,chiplets%20into%20a%20single%20semiconductor.
- https://www.cadence.com/content/dam/cadence-www/global/en_US/documents/tools/ic-package-design-analysis/chiplets-and-heterogeneous-packaging-are-changing-system-design-and-analysis.pdf
- https://www.chetanpatil.in/the-long-term-impact-of-semiconductor-chiplets/
- https://ieeexplore.ieee.org/document/9712583
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9499852&tag=1
- https://www.techpowerup.com/264271/amd-gives-itself-massive-cost-cutting-headroom-with-the-chiplet-design
- https://www.eetindia.co.in/chiplet-ecosystem-gathering-momentum/

MxM

- *https://www.wikiwand.com/en/Mobile_PCI_Express_Module#MXM_3.x_cards*
- *Mobile PCI Express Module Electromechanical Specification V 3.1*

NVLink

- *https://hc34.hotchips.org/assets/program/conference/day2/Network%20and%20Switches/NVSwitch%20HotChips%202022%20r5.pdf*
- *https://www.wikiwand.com/en/NVLink*
- *https://developer.nvidia.com/blog/how-nvlink-will-enable-faster-easier-multi-gpu-computing/*
- *https://graphicscardhub.com/mxm-graphics-card-gpu-list/*

NVSwitch

- *https://developer.nvidia.com/blog/nvswitch-leveraging-nvlink-to-maximum-effect/*
- *https://www.nextplatform.com/2018/04/04/inside-nvidias-nvswitch-gpu-interconnect/*

GPUDirect

- *https://developer.nvidia.com/blog/gpudirect-storage/*
- *https://docs.nvidia.com/gpudirect-storage/design-guide/index.html*

Miscellaneous 2.1

- *Steffora Mutschler, "Wrestling With High-Speed SerDes," Semiconductor Engineering, 2019. [Online]. Available: https://semiengineering.com/wrestling-with-high-speed-serdes/ .*
- *G. E. Moore, "Cramming more components onto integrated circuits. In: Electronics," Electronics, 1965.*
- *L. T. Su, S. Naffziger, and M. Papermaster, "Multi-chip technologies to unleash computing performance gains over the next decade," in Technical Digest - International Electron Devices Meeting, IEDM, 2018.*
- *T. Simonite, "To Keep Pace With Moore's Law, Chipmakers Turn to 'Chiplets,'" Wired, 2018. [Online]. Available: https://www.wired.com/story/keep-pace-moores-law-chipmakers-turn-chiplets/ .*
- *T. Coughlin, "Chiplets for All," Forbes, 2019. [Online]. Available: https://www.forbes.com/sites/tomcoughlin/2019/05/11/chiplets-for-all .*
- *E. Sperling, "Chiplets, Faster Interconnects, More Efficiency," Semiconductor Engineering, 2019. [Online]. Available: https://semiengineering.com/chiplets-faster-interconnects-and-more-efficiency/ .*
- *C. Sun et al., "Single-chip microprocessor that communicates directly using light," Nature, 2015.*
- *V. Stojanović et al., "Monolithic silicon-photonic platforms in state-of-the-art CMOS SOI processes [Invited]," Opt. Express, vol. 26, no. 10, p. 13106, 2018.*
- *M. Wade, M. Davenport, M. Rust, and F. Sedgwick, "TeraPHY : A Chiplet Technology for Low-Power , High-Bandwidth In-Package Optical I / O," Hot Chips Conference, 2019. [Online]. Available: https://ayarlabs.com/teraphy-a-chiplet-technology-for-low-power-high-bandwidth-in-package-optical-i-o/ .*
- *R. Danilak, "Why Energy Is A Big And Rapidly Growing Problem For Data Centers," Forbes, 2017. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2017/12/15/why-energy-is-a-big-and-rapidly-growing-problem-for-data-centers .*
- *J. Goergen, C. Systems, and V. Parthasarathy, "Spanning SERDES Across Reaches - Finding the Best Modulation Approach," no. November, 2014.*
- *R. Meade et al., "TeraPHY: A High-Density Electronic-Photonic Chiplet for Optical I/O from a Multi-Chip Module," 2019 Opt. Fiber Commun. Conf. Exhib. OFC 2019 - Proc., no. Mcm, pp. 5–7, 2019.*
- *D. Kehlet, "Accelerating Innovation Through A Standard Chiplet Interface : The Advanced Interface Bus (AIB)," 2019*
- *R. Mahajan et al., "Embedded Multi-die Interconnect Bridge (EMIB)-A High Density, High Bandwidth Packaging Interconnect," in Proceedings - Electronic Components and Technology Conference, 2016.*
- *M. Lapedus, "Bridges Vs. Interposers," Semiconductor Engineering, 2018. [Online]. Available: https://semiengineering.com/using-silicon-bridges-in-packages/ .*