nature electronics

Article

https://doi.org/10.1038/s41928-024-01126-y

High-performance, power-efficient three-dimensional system-in-package designs with universal chiplet interconnect express

Received: 24 July 2023

Accepted: 17 January 2024

Published online: 19 February 2024

Debendra Das Sharma [1], Gerald Pasdast [2], Sathya Tiagaraj [3] & Kemal Aygün [4]

Universal chiplet interconnect express (UCIe) is an open industry standard interconnect for a chiplet ecosystem in which chiplets from multiple suppliers can be packaged together. The UCIe 1.0 specification defines interoperability using standard and advanced packaging technologies with planar interconnects. Here we examine the development of UCIe as the bump interconnect pitches reduce with advances in packaging technologies for three-dimensional integration of chiplets. We report a die-to-die solution for the continuum of package bump pitches down to 1 µm, providing circuit architecture details and performance results. Our analysis suggests that—contrary to trends seen in traditional signalling interfaces—the most power-efficient performance for these architectures can be achieved by reducing the frequency as the bump pitch goes down. Our architectural approach provides power, performance and reliability characteristics approaching or exceeding that of a monolithic system-on-chip design as the bump pitch approaches 1 µm.

On-package integration of multiple dies has been widely deployed in commercial products. Examples include central processing units such as Intel's Sapphire Rapids[1] and AMD's EPYC and Ryzen[2], and general-purpose graphics processing units such as Intel's Ponte-Vecchio[3] and Nvidia's Hopper[4]. Tactically deployed for decades[5,6], this approach using chiplets—smaller chips that when packaged together realize the function of a larger chip—can meet ever-increasing compute demands and overcome die reticle limits and yield challenges in advanced process nodes. Rapid progress in advanced packaging technology[7–12] has allowed designers to interconnect chiplets on-package and keep Moore's law[13] going.

Beyond overcoming reticle limits and yield challenges, there are various additional compelling reasons for the use of on-package chiplets[14–17]. These include lower overall portfolio cost, where reuse of chiplets helps avoid intellectual property porting costs that are exploding with advanced process geometries[18], while also taking advantage of the lower costs associated with using more mature process nodes. There is also a time-to-market advantage as new chiplets can be added or substituted while reusing the old, which reduces design and validation cycles[14]. Bespoke solutions are also possible where system-in-package (SiP) variants are created by mixing chiplets with new capabilities with existing chiplets into a package. This provides an open plug-and-play infrastructure similar to PCI Express[19,20] and Compute Express Link (CXL)[21] offerings at the board level[14].

Universal chiplet interconnect express (UCIe)[15] is an open industry standard interconnect offering high-bandwidth, low-latency, power-efficient and cost-effective on-package connectivity between heterogeneous chiplets (Fig. 1a). Future three-dimensional (3D) packaging architectures are expected to scale bump pitches lower than the historical minimum values of around 90–110 µm for organic packages

[1]Intel Senior Fellow, Intel Corporation, Santa Clara, CA, USA. [2]Senior Principal Engineer, Intel Corporation, Santa Clara, CA, USA. [3]Staff Engineer, Intel Corporation, Santa Clara, CA, USA. [4]Intel Fellow, Intel Corporation, Chandler, AZ, USA. e-mail: Debendra.das.sharma@intel.com

 https://doi.org/10.1038/s41928-024-01126-y

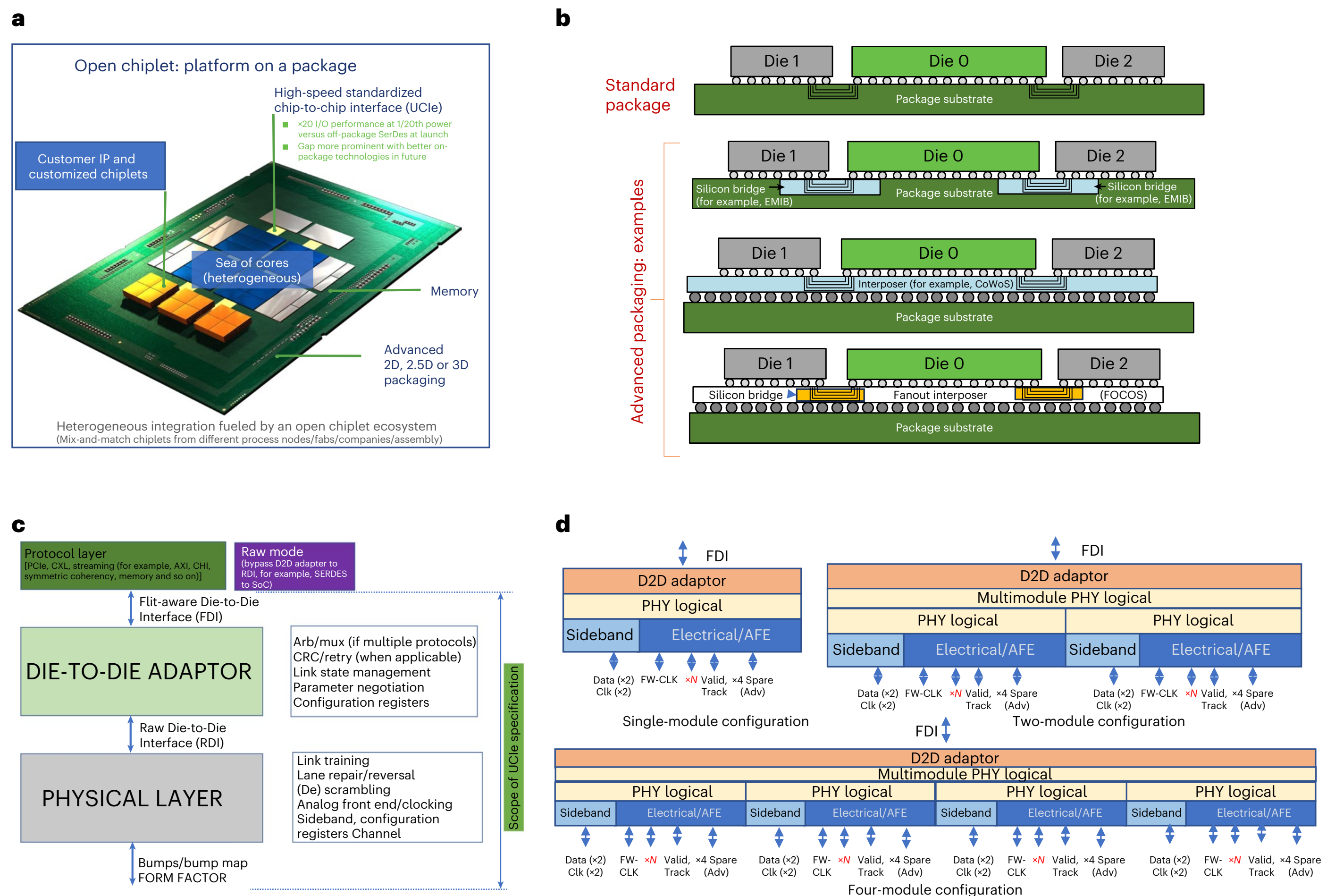

Fig. 1 | UCIe 1.0 specification application, physical interconnect, protocol stack and key metrics. a, Heterogeneous open chiplet on-package like that in a platform today—the open chiplet ecosystem UCIe supports mix and match of chiplets with different functionality across different foundries assembled in any assembly organization. **b**, UCIe 1.0 supports two types of packaging: standard with standard bumps and all flavours of advanced packaging with micro-bumps offered by different foundries and outsourced semiconductor assembly and test vendors, as shown by examples given in the figure showing 2D and 2.5D options. **c**, The layering approach of the UCIe 1.0 specification. **d**, Multimodule configurations applying UCIe. PCIe, peripheral component interconnect express; AXI, advanced extensible interface; CHI, coherent hub interface; RDI, raw die-to-die interface; SoC, system-on-chip; FDI, flit-aware die-to-die interface; Arb, arbitration; mux, multiplexer; EMIB, embedded multi-die interconnect bridge; COWOS, chip-on-wafer-on-substrate; FOCOS, fan out chip on substrate; AFE, analog front end; FW-CLK, forwarded clock.

and around 10–55 µm for enhanced two-dimensional (2D) architectures[18,22,23]. Recent work in this area has investigated the impact of various packaging options on the power, performance and area of the die-to-die (D2D) implementation of a set of reference system-on-chip devices[24].

In this Article, we consider how UCIe should evolve as bump pitches shrink in emerging advanced packaging architectures[11,12,22,23] and report a D2D solution for the continuum of package bump pitches down to 1 µm. For our approach, we provide the corresponding circuit architecture details, as well as detailed performance analysis. We also consider what needs to be done to enable use of our approach to construct SiPs of the future.

UCIe 1.0 specification overview

UCIe 1.0 defines two types of packaging (Fig. 1b): standard (UCIe-S) and advanced (UCIe-A). The standard package is used for cost-effective performance. The advanced packaging is used for power-efficient performance.

UCIe 1.0 is a layered protocol (Fig. 1c)[14–17]. The physical layer (PHY) is responsible for the electrical signalling, clocking, link training, sideband, circuit architecture and so on. UCIe supports different data rates, widths, bump pitches and channel reach to ensure the widest feasible interoperability, as detailed in Extended Data Table 1. The basic unit is a module comprising N single-ended, unidirectional, full-duplex data lanes (N = 16 for UCIe-S and N = 64 for UCIe-A), one single-ended lane for validation, one lane for tracking, a differential forwarded-clock per direction for the main band. The sideband consists of two single-ended lanes (one data and one 800 MHz forwarded-clock) per direction. The sideband interface is used for status exchange to facilitate link training, register access and diagnostics. Multiple modules (1, 2 or 4) can be aggregated to deliver more performance per link (Fig. 1d).

The D2D adaptor is responsible for reliable delivery of data through its cyclic redundancy check and link level retry mechanism[14–17]. When multiple protocols are supported, the adaptor defines the underlying arbitration mechanism. A 256-byte flow control unit defines the underlying transfer mechanism when the adaptor is responsible for reliable transfer. PCI Express and CXL protocols are mapped natively, as those are widely deployed at the board level across all segments of compute. Optical UCIe chiplets can transport CXL off-package to connect the processing and memory elements with high-bandwidth, low-latency and low-power connectivity to realize the vision of tightly coupled composable systems at the rack and pod level that are not

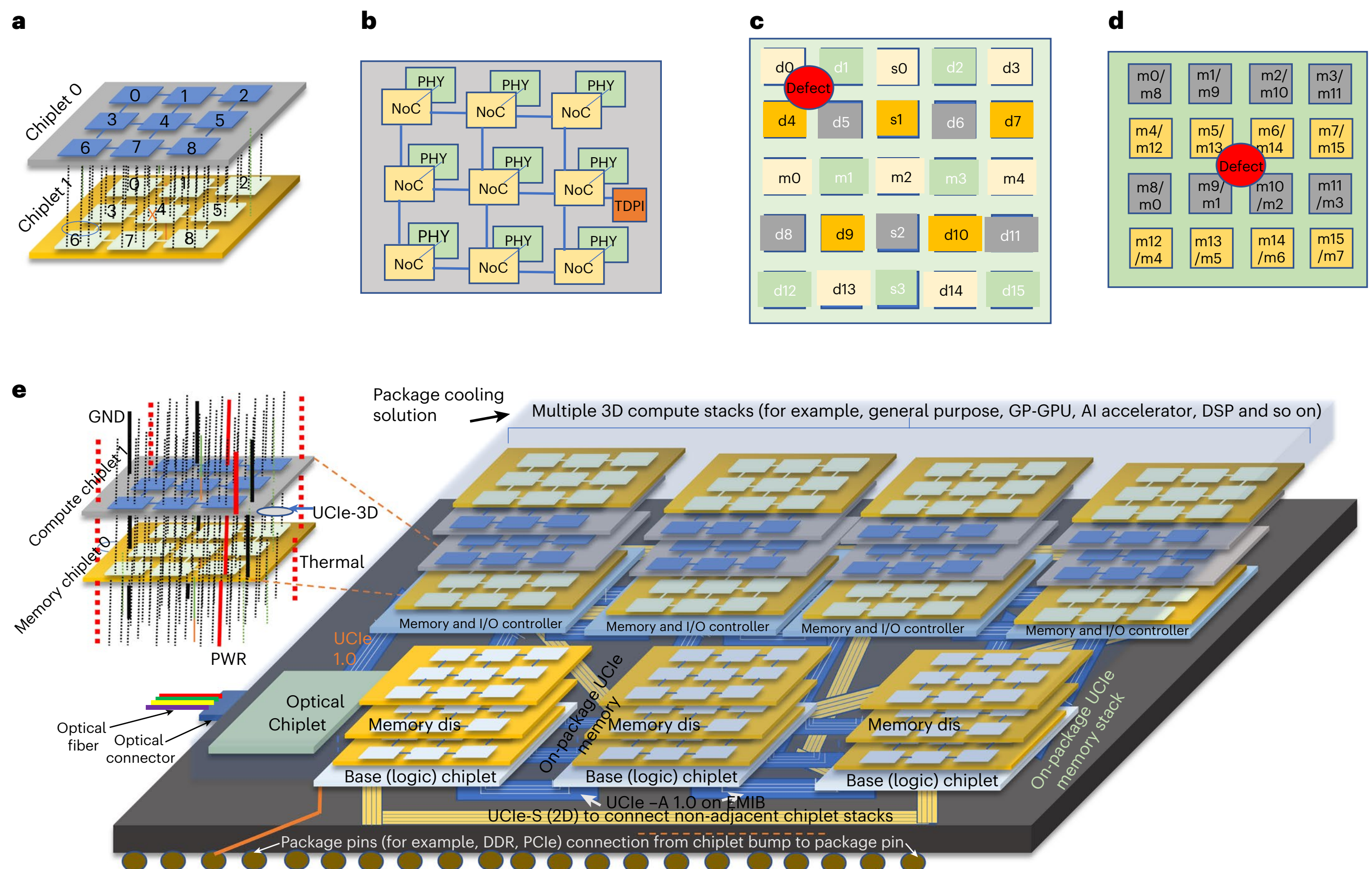

Fig. 2 | Chiplets connected with UCIe-3D. a, Two chiplets connected using UCIe-3D Links. The NoCs route traffic within and across chiplets. Any failure (NoC or UCIe-3D Link) can be routed around by other NoCs. **b**, Each NoC connects directly to one or more UCIe-3D hardened PHY block. Hardening means transistors are placed and routed for portability and optimal performance, power and area. Each chiplet has a common test, debug and pattern generation and checking infrastructure (TDPI) connected to one or more NoCs. This testing infrastructure is responsible for orchestrating training, testing and debug across the UCIe-3D Links by using the routing network of NoCs. As a result, the PHY does not have any configuration or status registers. The PHY is square and matches the size of the NoC to minimize any fan-in or fan-out of wires, so that the wire lengths are close to the least distance between NoC and PHY which will help minimize the area, power and latency. **c**, Each UCIe-3D link comprises an array of 25 subclusters, and each subcluster has 16 wires for a total of 400 wires; the 25 subclusters are data (d0–d15), miscellaneous (m0–m4) and spare (s0–s3). A defect (manufacturing, assembly, or run-time) can impact multiple adjacent subclusters. For getting around faults, the spares are connected as follows (where 'mux' indicates a many-to-one multiplexer): s0: mux{d0, d3, m0, m2, m4, d13, d14}, s1: mux{d4, d7, d9, d10}, s2: mux{d5, d6, d8, d11}, s3: mux{d1, d2, m1, m3, d12, d15}. This arrangement ensures that for any defect, up to four nearby submodules have a unique spare to use. Using the spares requires multiplexing of data and will result in additional gate-count. For the example defect shown in **c**, s0 will carry d0, s3 will carry d1, s1 will carry d4 and s2 will carry d5. **d**, An alternate implementation with 16 subclusters, each with 20+ wires, 16 of which are data and rest are miscellaneous (address, command, ECC and so on). In this arrangement, the NoC can optionally choose to degrade the link to half width (which is a 2:1 multiplex). **e**, Schematic illustrating the vision for a representative SiP system of the future based on the UCIe-3D architecture. EMIB, embedded multi-die interconnect bridge; GP-GPU, general purpose computing on graphics processing units; AI, artificial intelligence; DSP, digital signal processing; PWR, power; GND, ground.

possible with 2 m copper cables[14–16]. There is also provision to support other proprietary protocols as streaming protocols. UCIe also supports raw mode for applications such as connection to a high-speed serializer–deserializer (SERDES) die or a modem. In this case, raw bits are passed, bypassing the D2D adaptor by connecting to the raw D2D interface, as there is a complete protocol stack associated with that external interconnect on the main die side. Extended Data Table 1 summarizes the characteristics and the target performance metrics of UCIe 1.0.

Approach for UCIe-3D

In current applications, Chiplets are interconnected within a package both laterally (2D, 2.xD) and vertically (3D). Here, 2D refers to a standard organic packaging solution, while 2.xD refers to an advanced packaging solution with higher density connections. Some compute devices referenced earlier in this paper have chiplets laterally connected with 2D interconnects[6] or 2.xD (ref. 1). Memory device manufacturers have used 3D interconnected chiplets for more than a decade[25,26]. A combination of both 2.xD and 3D examples also exist[6], combining the advantages of both lateral and vertical interconnect. One recent key trend—especially for 3D packaging technologies, such as hybrid bonding (HB)—has been the aggressive shrinking of the bump pitches between the chiplets and the consequent reduction of the corresponding interconnect distances and their associated electrical parasitics[11,12,22,23]. As bump pitch decreases, the area under the bump reduces, and the number of wires for a given area increases as a square of the bump pitch reduction. With orders of magnitude in wire density increase and area reduction, an architectural approach completely different from UCIe 1.0 should be pursued. When architected correctly, as demonstrated with our approach, interconnected chiplets with these low bump pitches will offer better latency and power characteristics than large monolithic dies and will offer the same benefits that Moore's law has provided with reduced transistor sizes for over 50 years.

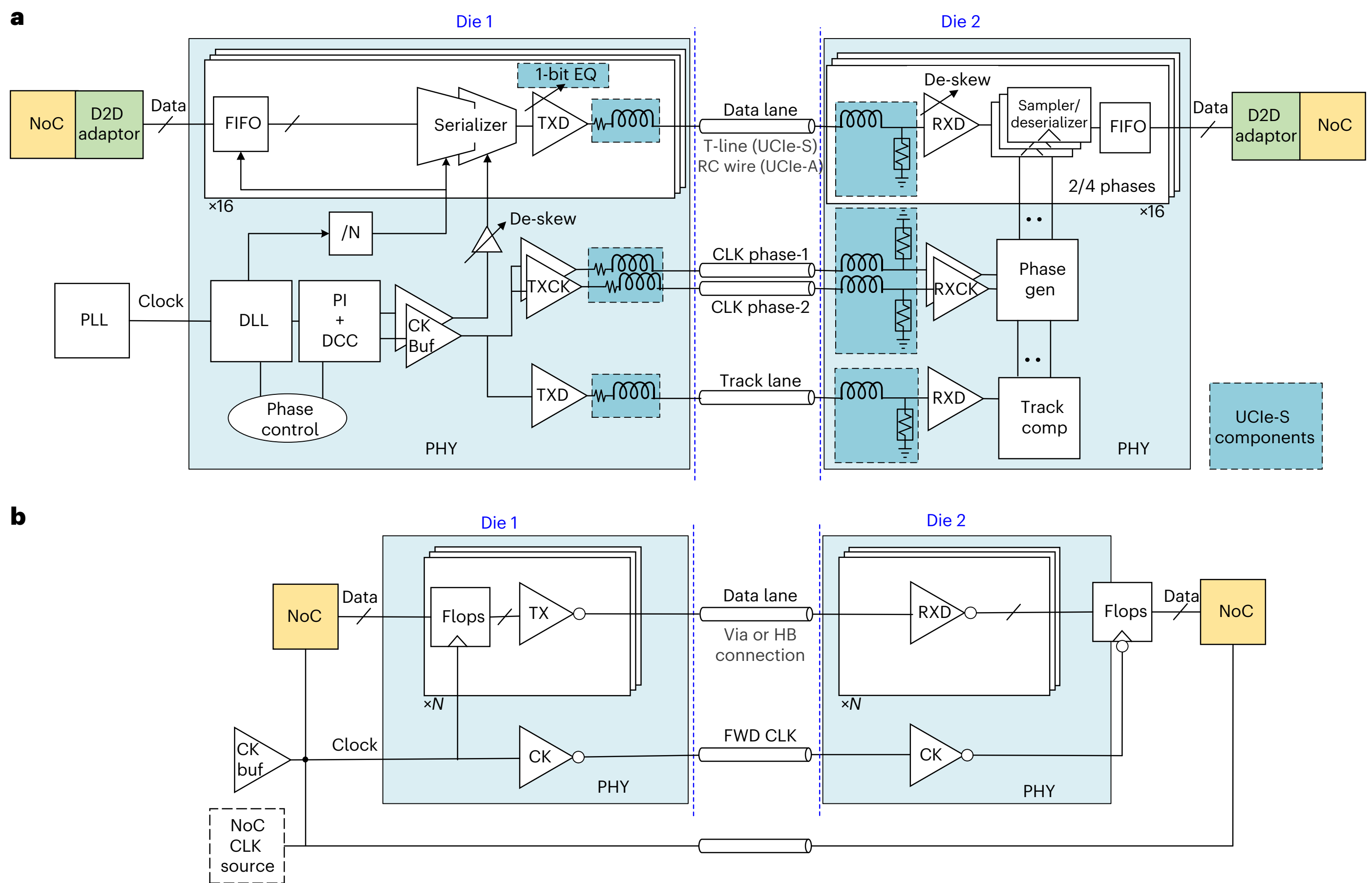

Fig. 3 | Circuit architecture of UCIe. a, UCIe-S and UCIe-A interconnect architecture for 2D and 2.xD, respectively. The removal of components for UCIe-A is indicated in dotted boxes. The UCIe 1.0 PHY is based on a 'matched' source synchronous clocking architecture with most components composed of high-speed 'digital' (that is, complementary metal oxide semiconductor switching) circuits including the delay locked loop, phase interpolator, duty cycle connector and TX driver. The RX implementation can be a simple inverter at lower speeds and requires a more sensitive architecture at higher speeds. Common high-speed PHY features such as clock-to-data centring, lane-to-lane de-skew, TX driver impedance are needed for both UCIe-S and UCIe-A PHYs at higher speed operation. In addition, due to the transmission line nature of 2D interconnect channels, UCIe-S also requires TX and RX equalization as well as inductor coils at the TX and RX pads for pad capacitance reduction at higher speeds to manage intersymbol interference. UCIe-A's elimination of area-intensive inductors, resistors and the equalization are essential in allowing for area compaction to support the tighter bump-pitch range. This translates to substantial KPI improvements over UCIe-S including ×6 linear bandwidth density, ×11 areal bandwidth density and ×2 improved power. **b**, The proposed PHY architecture for UCIe-3D based on our implementation is dramatically simplified for the <10 µm bump pitch support target range. Maximum data rate is proposed to be 4 GT s^{-1}, which should cover most on-die logical speeds in use today and for the foreseeable future. We suggest source-synchronous clocking across the PHY with simple setup and hold specifications at the bumps to enable static-timing validation. Boundary flops to retime NoC output data signals before the TX will minimize lane-to-lane skew at the receiver output. Data rate, distance to NoC, process and timing will determine whether a retiming flop is needed between the receiver output and NoC input, and hence the retiming flop in the diagram is shown straddling the PHY boundary and is implementation specific. The NoC clock source can reside on either die. We propose a 0.01 pJ b^{-1} target to enable implementations equal to or better than full-monolithic implementations. FIFO, first-in-first-out; CK, clock; DLL, delay locked loop; PI, phase interpolator; DCC, duty cycle corrector; Buf, buffer; TXCL, transmitter clock; TXD, transmitter data; /N, divide by N; PLL, phase locked loop; EQ, equalization; T-line, transmission line; RC, resistance-capacitance; CLK, clock; RXCK, receiver clock; Gen, generator; Comp, compensation.

We suggest the next generation of UCIe will continue to be unidirectional, supporting both 2.xD and 3D connectivity, running at the chiplet internal frequency or even a lower frequency. We refer to our proposed next generation of UCIe as UCIe-3D. Our approach will result in orders of magnitude improvement in bandwidth and power efficiency over the existing UCIe 1.0 specification. The lower frequency and short distance make the circuits simpler; they will fit within the bump area and operate with lower power. Since this interconnect will have a lower bit error rate (BER) due to the short distance and low frequency, we propose to remove the D2D adaptor completely.

Our vision is that two chiplets will connect using multiple independent modules, with each UCIe-3D PHY directly controlled by the Network-on-Chip controller (NoC) (Fig. 2a). The common functionality across all PHYs is orchestrated by a common control block in the chiplet to amortize the overhead (Fig. 2b). The PHY is implemented using a square bump layout with dedicated subclusters for data versus non-data (address, error correction code (ECC), spares and so on). Repairing around defects is managed at the NoC and chiplet level (Fig. 2a–d).

The PHY architecture for UCIe-S and UCIe-A is based on a forwarded-clock (source synchronous), parallel input–output (IO) structure with most of the building blocks constituted as high-speed complementary-metal-oxide semiconductor circuits. A typical speed for the logical interface to the PHY runs at 2 GHz (ref. 15). Thus, at 32 GT s^{-1} operation, there is a 16:1 serialization and 1:16 deserialization (SERDES) factor between the logical interface and the PHY transmitter/

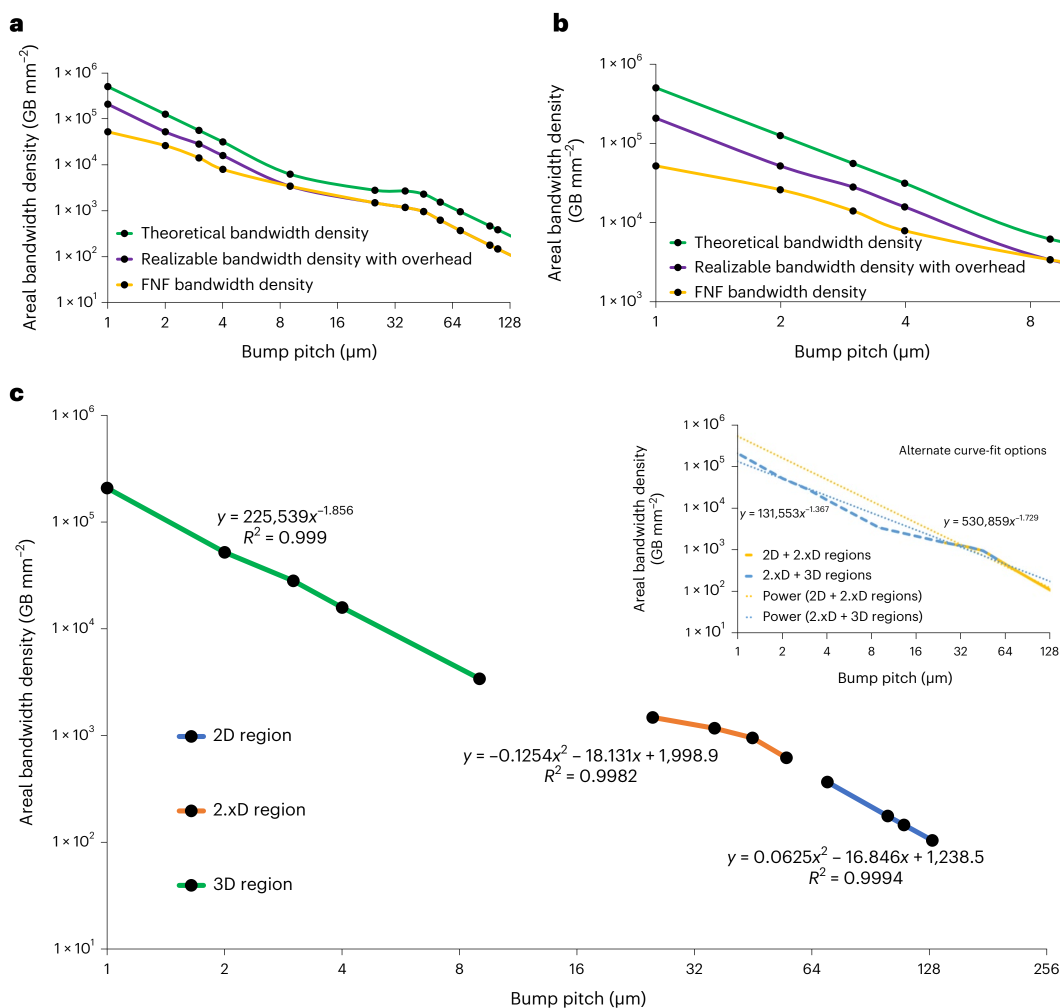

Fig. 4 | Areal bandwidth density comparison. a,**b**, Theoretical, realizable and FNF bandwidth densities versus bump pitch for the full range of bump pitches (**a**) and for bump pitches <9 μm (**b**). Theoretical bandwidth density values are derived using equation (2) with no overheads and an assumption of maximum frequency per pin, as stated by the UCIe specification. Realizable and FNF density values are calculated using equation (3). In **a**, the UCIe specification has been followed till 25 μm pitch (32 GT s^{-1} down to 45 μm, 24 GT s^{-1} at 36 μm and 12 GT s^{-1} at 25 μm). In **b**, for 9 μm bump pitch and below, the parameters used have been described above. **c**, Demonstration of the 'practical' approach to analysis. As explained in the Methods section, the inference is that the error can be as high as ×10 (one order of magnitude) if we curve-fit all the bump pitches to one or two equations, as shown in the inset of **c**. A more practical approach is to split the equation for projections into three regions, as shown in **c**. One equation governs the 2D region (130 μm down to 90 μm), another the 2.xD region (ranging from 65 μm to 25 μm) and a final one the 3D region (ranging from 16 μm down to 1 μm). The 2D region is almost linear, with little weighting to a quadratic region; the 2.xD region is primarily quadratic; and the 3D region is dominated by power series. This approach limits the prediction error to be less than 8% across a wide range of bump pitches.

receiver (TX/RX) operation across the package interconnect. Running at sufficiently high SERDES factors is critical when packaging interconnect is practically limited by layer count and IO bump pitch as more data needs to travel across each package wire. As the bump pitch decreases, for example, from 110 μm for UCIe-S to 45 μm for UCIe-A, a substantial increase in D2D bandwidth as well as silicon area bandwidth density is realized. For UCIe-3D, this bandwidth trend further accelerates to a point where it is more efficient to run the IO at the native NoC frequency without the need for any SERDES.

We suggest the elimination of the D2D adaptor and simply having the NoC directly interface with the UCIe-3D circuits. The NoC designer will set the supply voltage level to the appropriate value to meet the needs of the NoC logic timing. The most efficient UCIe-3D interconnect will be one that can operate on the same supply as the NoC to avoid any special supply requirements. We suggest a lean D2D data path that consists simply of a retiming flop stage at the UCIe-3D TX bump followed by an appropriately sized inverter driver to meet its own up-to-5 V charged-device model (CDM) electrostatic discharge (ESD) requirement (via parasitic diodes), as well as the slew rate requirements across the HB connection into the RX inverter and ESD on the other die. We anticipate a move to 0 V CDM requirement as bump pitches get down to 3 μm so that the UCIe-3D PHY fits within the bump area. Figure 3 and Extended Data Table 2 show the evolution of UCIe PHY architecture from the -S and -A variants to the UCIe-3D solution.

The UCIe-3D approach is amenable to synthesis and automatic place-and-route tools and adaptable to a wide range of floor plans. It will be highly desirable to enable static timing analysis for timing closure for the D2D crossing, and to facilitate that we suggest specifying timing at the HB bump boundary and continuing with the forwarded-clock architecture of UCIe-S and UCIe-A to establish a set of clock-to-data specifications at bump pins.

Since the same architecture is used across both sides of the 3D connection, asymmetric bandwidth needs can be addressed by simply arraying different number of IO modules for each side of the

 https://doi.org/10.1038/s41928-024-01126-y

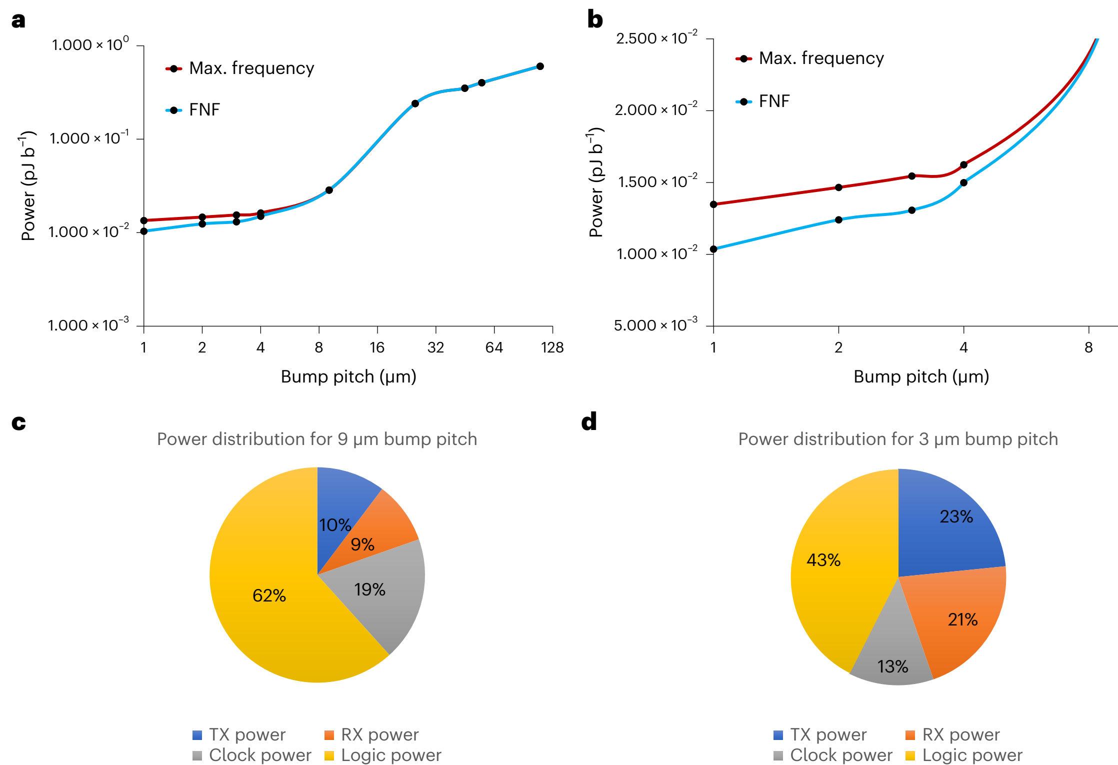

Fig. 5 | Power efficiency comparison. **a**, Trend of D2D power versus bump pitch at maximum frequency of operation. **b**, The benefit of FNF link at sub-9 µm pitches. The curves in **a** and **b** assume maximum frequency as specified in UCIe specification till a certain bump pitch (32 GT s^{-1} till 45 µm, 24 GT s^{-1} at 36 µm and 12 GT s^{-1} at 25 µm). At 9 µm, the total power of the D2D link due to simplified architecture at 4 GT s^{-1} is 0.03 pJ b^{-1}. At 1 µm, the total power drops to 0.015 pJ b^{-1} at 4 GT s^{-1}. Use of FNF at 1 GT s^{-1} enables an additional 50% power savings; for example, at 1 µm, the total power is 0.01 pJ b^{-1}. The UCIe-3D link is primarily dominated by leakage in the TDPI at smaller bump pitches, assuming one TDPI per 512 full-duplex data lanes. With further disaggregation and NoC-based data path controls, there is an opportunity for reducing the power further. **c**,**d**, Example power distribution of the UCIe-3D link at 9 µm (**c**) and 3 µm (**d**), respectively. The transmit power component is dominated by the capacitance and remains almost identical since the scaling of bump parasitics in 3D is negligible as bump pitch shrinks. The 5 V ESD adds 15–24% of the total power depending on the pitch. For 30 V CDM, the contribution of ESD to D2D power will be higher. Future trends of the ESD, where CDM levels are projected to be smaller[30], will help mitigate this effect. Clock power also includes the forwarded-clock power required for sending to the other die. Logic power scales as a function of area, and the total number of inverters and logic reduces as pitch shrinks, reducing its power share from 62% to 43%. As explained above, since TDPI remains nearly unchanged as pitch shrinks, the power at 3 µm and a portion of power at 9 µm is still leakage dominated. Max., maximum.

connection. The TX, RX and clock circuits are simple inverters, creating a matched data and clock path with data launched at rising clock edge and captured with the corresponding forwarded falling clock edge. The forwarded-clock source is the same as the NoC clock source and is shared on both dies to avoid the power and latency issues associated with clock domain crossings. At bump pitches approaching 3 µm and below, we expect that a fractional NoC frequency (FNF) D2D crossing may be advantageous for power optimization. For example, a D2D crossing at 1 µm bump pitch running at a native NoC frequency of 4 GHz may consume more power than running twice the number of wires at 2 GHz. Standard loopback schemes such as near end (within die) or far end (at the D2D crossing) will need to be incorporated into the overall data path to enable detection of defects at sort testing before assembling multiple dies within a package.

Analysis of UCIe-3D

We next analyse the efficiency of the UCIe-3D approach and present the results from our implementation on an Intel process node (see Methods for further details).

As bump pitch scales, the theoretical bandwidth density can be calculated using these formulae:

$$\text{Bump density} = \frac{1}{\text{Pitch}^2} \quad (1)$$

$$\text{Theoretical bandwidth density} = \text{Bump density} \times \text{Data rate} \quad (2)$$

For example, for pitch = 9 µm, bump density = 12,346 bumps mm^{-2} and assuming data rate = 4 GT s^{-1}, theoretical bandwidth density = 6,173 GB s^{-1} mm^{-2}. Considering the practical overheads, power and ground bumps, and bump efficiency due to a hex versus square bump pattern, equation (2) becomes

$$\text{Realizable bandwidth density} = \left(\frac{1}{\text{Pitch}^2}\right) \times \text{Bump efficiency} \times (1 - \text{Data, repair, power and ground overhead}) \times \text{Data rate} \quad (3)$$

where

$$\text{Bump efficiency} = \begin{cases} 1.15 \text{ for hex} \\ 1 \text{ for square} \end{cases} \quad (4)$$

$$\text{Power ground overhead} = \begin{cases} 0.35 \text{ for } 9\,\text{um} \le \text{pitch} \le 130\,\text{um} \\ 0.4 \text{ for } 2\,\text{um} \le \text{pitch} \le 9\,\text{um} \\ 0.5 \text{ for pitch} < 2\,\text{um} \end{cases} \quad (5)$$

https://doi.org/10.1038/s41928-024-01126-y

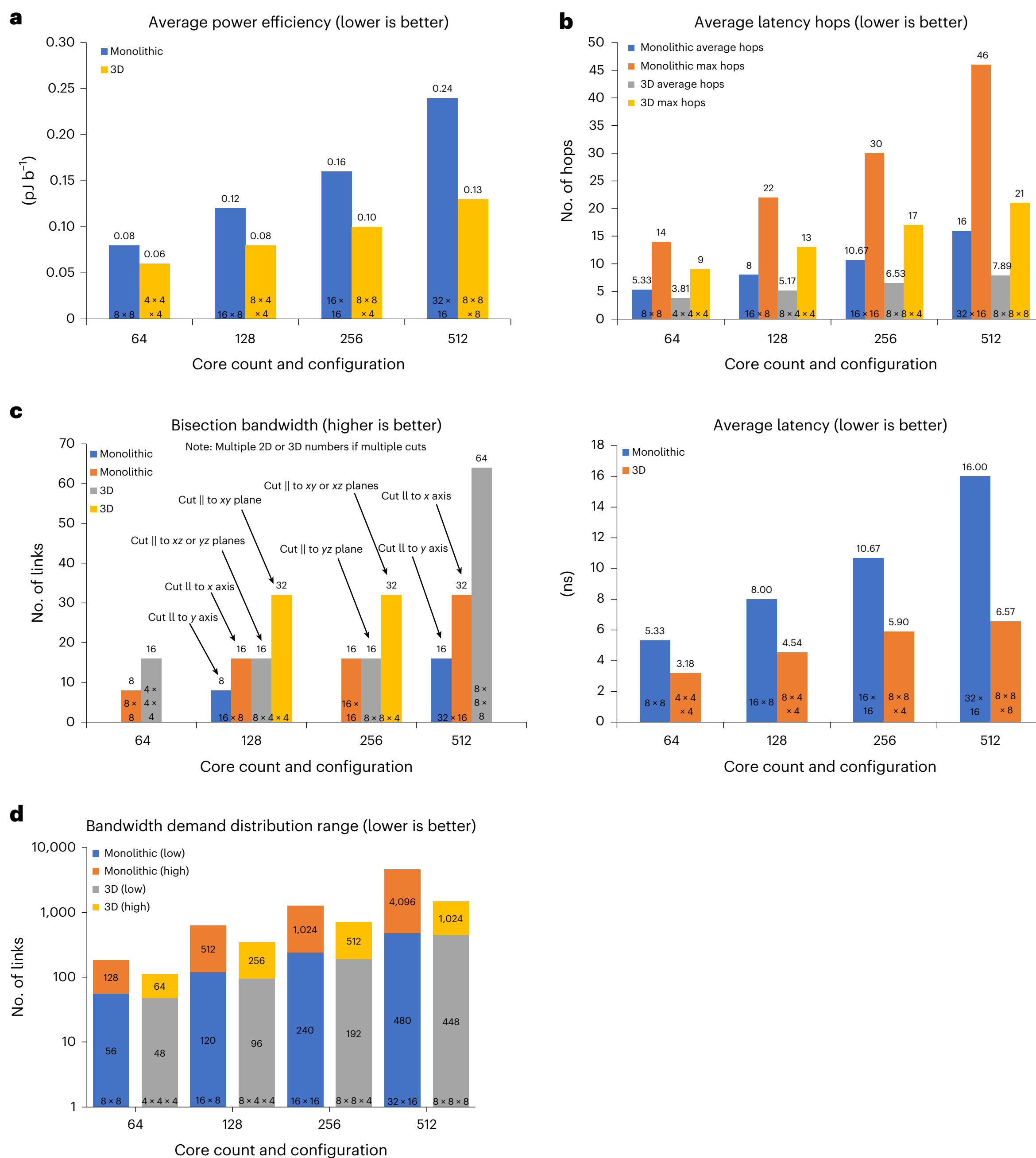

Fig. 6 | Power and performance comparison between UCIe-3D and monolithic. a, Reduction in average power with the proposed UCIe-3D as compared to a monolithic solution, due to lower number of hops and shorter travel distances. The comparison is over four core-count scenarios (64, 128, 256 and 512 cores) as shown along the graph's *x* axis; physical configurations are listed inset in the bars (as an example, '8 × 8', refers to an 8 × 8 array of cores on a planar monolithic die, and '8 × 8 × 8' refers to 8 dies, each 8 × 8, stacked using UCIe-3D). The power values are determined from equation (9), based on the 3 mm distance and on equation (10) using 0.02 pJ b^{-1}. The trends would be similar if the latency number or power number varied between hops in monolithic dies. **b**, Reduction of average latency, average hops and maximum number of hops with UCIe-3D as compared to a monolithic solution, for the same set of scenarios as in **a**. The 3D hop count is determined by equation (8), based on the 150 µm distance between the NoCs plus the optional FIFO latency. **c**, Left, increase in bisection bandwidth with UCIe-3D as compared with a monolithic solution, due to extra vertical links and right, reduction in average latency, in the scenarios of **a** and **b**. **d**, Reduction in bandwidth demand distribution and the maximum bandwidth demand for any link with UCIe-3D as compared to a monolithic solution, in the scenarios of **a**, **b** and **c**. The bandwidth demand distribution represents the congestion in the links with a higher number indicating higher congestion. As expected, we see higher bandwidth demand distribution increasing towards the middle row, column and vertical links, since most of the communication goes through those links. A lower number means better distribution, which reduces the congestion.

Here, the overhead for data bumps is assumed to be ~3% (16 lanes for a given 512-bit bus (TX + RX)) to accommodate sideband, clocks, track and valid signals. Hex pattern has been assumed for 2D and 2.xD, while for fine pitch bonding, a square pattern is assumed. The repair is managed at the network and cluster levels, as explained previously. Our estimate for repair overhead is 10%, given the multiple link connectivity across two dies. For comparison, the repair overhead today for 2.xD is around 3%. The Power ground overhead in equation (5) remains consistent even at tighter bump pitches. At higher bump pitches, the maximum data rate is higher, which forces the need for

additional ground bumps for isolation and adequate signal integrity. At lower pitches, the current per bump is limited due to drawn dimensions, necessitating additional power and ground bumps for robust power delivery. It is well understood that as interconnect technologies continue to evolve, the need for additional power bumps can diminish and help increase achievable bandwidth density. Theoretical, realizable and FNF bandwidth densities with these overhead assumptions and for bump pitches from 128 μm to 1 μm have been plotted in Fig. 4. These plots show that the theoretical bandwidth density at 9 μm pitch is same as the realizable bandwidth density at 3 μm pitch or the FNF bandwidth density at 2 μm, showing the impact of the various overheads and the need for continued improvement in interconnect technologies to achieve tighter bump pitches. Curve fitting the points from the above equations results in the following equation to predict the realistic bandwidth density for any bump pitch with realistic overheads:

$$\text{Realizable bandwidth density} \approx \begin{cases} 0.0625x^2 - 16.846x + 1,238.8; 90\,\mu m \leq \text{ pitch} \leq 130\,\mu m \\ -0.1254x^2 - 18.131x + 1,998.9; 25\,\mu m \leq \text{ pitch} \leq 65\,\mu m \\ 225,539x^{-1.856}; 1\,\mu m \leq \text{ pitch} \leq 16\,\mu m \end{cases} \quad (6)$$

FNF enables us to disaggregate system-on-chip devices further without crossing the thermal limits of the SiP and helping the key performance indicators (KPIs), including power-performance-area of the product. At 9 μm, the max frequency of 4 GT s^{-1} has been assumed (in accordance with the UCIe-3D approach) and in Fig. 4a,b. However, with FNF, we will limit max speed to 2 GT s^{-1} from <9 μm to 2 μm and 1 GT s^{-1} at 1 μm. This will still enable ×2 bandwidth density increase as pitch goes from 9 μm to 1 μm, while saving a sizeable portion of power, as explained in the power section below.

Equation (6) has good correlation and fit to the practical implementations we have designed and helps extrapolate bandwidth density to a wide range of pitches that can be used by UCIe-S, UCIe-A and UCIe-3D. Enough overlaps have been given in the range considered for curve fitting to keep the boundaries between 2D, 2.xD and 3D interconnect regions flexible.

With disaggregation spanning around multiple cuts along multiple axes, minute changes in D2D power quickly add up at the system level. A D2D link complicates the timing between two dies and hence increases the need for additional complexity of the IO and testing, as explained in earlier sections. Where the two dies can be completely different process or material, or even the same process but operated at different voltages, optimized for performance, the total number of components in the data path increases. Additional complexities—including standalone die testing and yield recovery of defects using lane repair—will also add more components on the data path. In addition to these, ESD adds a notable component, a capacitor, to the D2D crossing. Figure 5 shows the D2D power as a function of bump pitch (for 2D, 2.xD and 3D regions) and the power distribution of the UCIe-3D link at 9 μm and 3 μm bump pitches.

With the UCIe-3D (that is, quasi-monolithic) architecture, the latency adder can be a couple of flops on both sides of the die, allowing maximum flexibility for the two dies to be on two different processes or voltages. With adequate tool development, it is also possible to converge timing across two dies and reduce the latency further to look like monolithic IO. As the bump pitch shrinks, the distance from the controller to the bump reduces, enabling us to remove stage flops and make the 3D crossing look monolithic. Besides the lower circuit and logic latency, there are architectural performance benefits with 3D stacking of dies. Overall hop latency is reduced and provides higher bandwidth, compared to the planar arrangement (monolithic, 2D or 2.xD interconnects). Figure 6 summarizes the performance for compute elements implemented as an ideal monolithic die (with no yield or reticle constraints) versus chiplets connected using UCIe-3D, both using mesh topology with identical bandwidth per hop. We use an ideal monolithic die for KPI performance comparison, since it outperforms UCIe 1.0-based chiplet designs. Similar trends will exist for other applications, such as memory or a combination of compute elements and memory.

Finally, for reliability, we use failure in time (FIT), which is the number of failures in 10^9 hours. Ideally, a chiplet should have a FIT $\ll 1$ for all its UCIe links combined so that the contribution of the links is an exceedingly small fraction of the typical FIT in a chiplet (100's). While we expect the chiplet to have error detection and correction that goes unchanged in the UCIe links, we have ignored any form of ECC and that all lanes in all the links are active at the same time to obtain a pessimistic estimate of FIT. We suggest that the specified BER be 10^{-30}; this will result in a FIT of 3.6×10^{-4} for 100 Tb s^{-1} bandwidth (Methods). With the built-in error detection and correction logic in each chiplet, this number will be orders of magnitude lower, even for a target of 10^{-27} BER, as demonstrated subsequently with an example.

Conclusions

We have reported UCIe-3D, a power-performance efficient and cost-effective method for constructing SiP architectures using emerging advanced 3D packaging technologies with shrinking bump pitches. UCIe-3D offers superior performance with lower latency, higher bisection bandwidth and lower bandwidth demand than planar implementations of 2D and 2.xD interconnects or even a large monolithic die.

Our approach could be used to create powerful SiPs. In particular, the UCIe-3D approach could be used to create multiple 3D heterogeneous compute stacks—each with their own local memory chiplets, multiple on-package memory stacks and external I/O and memory chiplets—all connected internally across and through existing UCIe 1.0 interconnects. In this architecture, each chiplet can be connected to the chiplet above or below in a face-to-face, face-to-back, back-to-face, or back-to-back configuration[27]. In non-face-to-face connection scenarios, signals would need to travel through silicon vias. Further exploration is required into the development of silicon-via manufacturing and assembly technologies that can scale with the bump-pitch range and introduce negligible electrical parasitics while preserving the KPIs at current levels.

Additional challenges are also likely in the areas of cooling, power delivery and reliability. The thermal demands of emerging 2.5D and 3D packaging architectures in terms of an increase of the hotspot peak power densities compared to average values have already been highlighted[18]. Additional 3D-stacked chiplets will only exacerbate this problem and will require more advanced cooling capabilities. The power delivery for such an architecture is also expected to create new issues, and an increased reliance on technologies such as backside power delivery is likely[28,29]. For reliability, there will be a need for repair strategies, along with reduced electrostatic discharge protection requirements for the assembly processes[30].

Finally, advances in electronic design automation are necessary. The need for such design automation capabilities has been highlighted previously[27], and further innovations[31] are required to create true mix and match architectures.

Methods

Reliability calculation

We will first calculate the FIT assuming no error correction or detection. We also assume that all the UCIe links are running at full bandwidth all the time. Thus, any bit error over time would be considered fatal and potentially cause silent data corruption (SDC). The FIT here denotes the SDC component only, since there is no detection.

Let p_{ber} denote the probability of a bit error and b denote the total bandwidth demand in terabits per second across all UCIe links for a given chiplet. Let n denote the number of bits in 10^9 hours. Thus,

$$n = 3,600 \times 10^9 \times b \times 10^{12} = 36 \times b \times 10^{23} \quad (7)$$

Since FIT is the probability that no failures happen over 10^9 hours, and given that we are not taking advantage of any error detection or correction, we can express FIT as:

$$\mathrm{FIT} = 1 - (1 - p_{\mathrm{ber}})^n \approx n \times p_{\mathrm{ber}} = 36 \times b \times p_{\mathrm{ber}} \times 10^{23} \quad (8)$$

For 100 Tb s^{-1} bandwidth, $b = 100$. Thus, from equation (8), with a p_{ber} of 10^{-30}, we obtain a FIT of 3.6×10^{-4}.

Next, we provide an example assuming that the chiplets have an inherent (137, 128) single error correct, double error detect code that will correct all single bit flip and detect all double bit flips in the code word of length 137 bits. Let c denote the number of codewords in 10^9 hours:

$$c = \frac{n}{137} \quad (9)$$

The probability that a code word will be received correctly is the probability of 0 or 1 errors in the code word. Any even errors will be detected, which are categorized as detected uncorrected error (DUE). Any odd number of errors, 3 or more, has a chance of aliasing to either a valid code word or correcting a bit and potentially causing an SDC event. For simplicity, given the low BER, we will calculate the probability of a code word causing DUE to be the probability of 2 errors ($p_{\text{code-due}}$) and a code word causing SDC to be the probability of 3 errors ($p_{\text{code-sdc}}$), as follows:

$$p_{\text{code-due}} \approx \binom{137}{2} (1 - p_{\mathrm{ber}})^{135} p_{\mathrm{ber}}^2 \approx 9316 \times p_{\mathrm{ber}}^2 \quad (10)$$

$$p_{\text{code-sdc}} \approx \binom{137}{3} (1 - p_{\mathrm{ber}})^{134} p_{\mathrm{ber}}^3 \approx 4.2 \times 10^5 \times p_{\mathrm{ber}}^3 \quad (11)$$

$$\mathrm{FIT(SDC)} = 1 - (1 - p_{\text{code-sdc}})^c \approx c \times p_{\text{code-sdc}} \quad (12)$$

$$\mathrm{FIT(DUE)} = 1 - (1 - p_{\text{code-due}})^c \approx c \times p_{\text{code-due}} \quad (13)$$

If p_{ber} is 10^{-30} and b is 100 Tb s^{-1}, FIT(SDC) = 1.1×10^{-50} and FIT(DUE) = 2.45×10^{-20}; if p_{ber} is 10^{-27} and b is 100 Tb s^{-1}, FIT(SDC) = 1.1×10^{-41} and FIT(DUE) = 2.45×10^{-14}, which is still acceptable.

Performance calculations

We will do the calculations assuming an $n_x \times n_y$ 2D mesh and an $n_x \times n_y \times n_z$ 3D mesh. If n is the total number of nodes, then $n = n_x \times n_y$ for the 2D mesh and $n = n_x \times n_y \times n_z$ for the 3D mesh topology. The 2D mesh is a special case of 3D mesh, where $n_z = 1$. The maximum number of hops is $(n_x - 1) + (n_y - 1) + (n_z - 1) = (n_x + n_y + n_z - 3)$, between nodes $\{0, 0, 0\}$ and $\{(n_x - 1), (n_y - 1), (n_z - 1)\}$.

The average distance (in hops or in nanoseconds) or power between any source–destination pair in a 2D or 3D mesh can be obtained by summing the distance between all source–destination pairs $\{i, j, k\}$ and $\{a, b, c\}$ multiplied with the appropriate weight (w_x, w_y, w_z) and dividing by the total number of source–destination pairs $n \times (n - 1)$ as follows:

$$\frac{\sum_{i=0}^{n_x-1} \sum_{j=0}^{n_y-1} \sum_{k=0}^{n_z-1} \sum_{a=0}^{n_x-1} \sum_{b=0}^{n_y-1} \sum_{c=0}^{n_z-1} (|i-a| \times w_x + |j-b| \times w_y + |k-c| \times w_z)}{n(n-1)} \quad (14)$$

The bisection bandwidth is the number of links for any cross-section with half the number of nodes on either side. For planes parallel to the x–y, y–z and x–z planes, the number of links are $n_x n_y$, $n_y n_z$ and $n_x n_z$, respectively.

Bandwidth distribution across links

For the routing in the mesh, to ensure deadlock freedom, the message traverses in the shortest x-distance first, then the shortest y distance, followed by the shortest z distance. Thus, the path from node (0, 1, 1) to node (3, 3, 3) is (0, 1, 1) → (1, 1, 1) → (2, 1, 1) → (3, 1, 1) → (3, 2, 1) → (3, 3, 1) → (3, 3, 2) → (3, 3, 3). For a node with 6 neighbours in a 3D mesh, there are two links each on the x, y and z directions, one connecting to a node that is 1 greater in that dimension and the other that is 1 lower in that dimension. We derive the equations for the increasing dimensional links, and the respective decreasing dimensional links are identical. For any node (a, b, c) in the increasing x dimension, traffic comes from $(a + 1)$ nodes (that is, nodes $(0, b, c)$, $(1, b, c)$, … (a, b, c)) that route through the link (a, b, c) to $(a + 1, b, c)$, denoted as L_x^a, targeting any node (x, y, z) where $x > a$, since the x dimension is routed first. That means we have $(n_x - a - 1) \times n_y \times n_z$ destination nodes that are routed through L_x^a. Thus, node (a, b, c) in its increasing x dimension gets traffic between $(a + 1) \times (n_x - a - 1) \times n_y \times n_z$ source–destination nodes. Similarly, in the increasing y and z dimensions of node (a, b, c) we get traffic between $(b + 1) \times n_x \times (n_y - b - 1) \times n_z$ and $(c + 1) \times n_x \times n_y \times (n_z - c - 1)$ source–destination nodes, respectively. The results reported in Fig. 6 are based on computing using this methodology for the bandwidth distribution.

Methods for leakage reduction and active power reduction

As bump pitch shrinks, the distance between the drivers gets reduced, resulting in a reduction of the total number of repeaters. Going from 9 µm to 3 µm, for the same shoreline and bandwidth targets, the total depth reduces by the square of the pitch. Aligning the shoreline to the controller reduces latency and the need for fan-out of signals. For example, with a shoreline of 1 mm and bandwidth of 800 GB s^{-1}, the depth of the PHY with 50% overhead on power and ground is ~350 µm. At 3 µm bump pitch, the depth of the PHY reduces to 50 µm for the same shoreline. At 50 µm depth, the total number of repeaters and flops needed reduces drastically. Clock distribution power reduces by area scaling, reducing the number of stages of clock distribution for the same number of bits. We also observed that the TX power reduces, since the bump dimensions get reduced from 9 µm to 3 µm HB.

Referencing the external UCIe specification, the total power contribution due to D2D (with all the overheads and logic needed for standalone testing) is 0.25 pJ b^{-1} at 0.5 V running at 16 GT s^{-1}. Since the 3D link is a much simpler interface compared to D2D IO operating at 16 or 32 GT s^{-1}, the VCC can be substantially lowered to help power savings. The power savings due to voltage reduction is substantial, irrespective of the inherent differences due to the design, underlying process technology, and variations due to process, voltage, and temperature, in the power calculation, due to the VCC^2 scaling factor.

Method for bandwidth density curve fitting

A detailed examination of the curve in Fig. 4 reveals that one curve does not fit all the spectrum of bandwidth densities from 130 µm down to 1 µm. The reason for this is that the repair ratio is different and the power ground overhead is different, and hence the equations cannot predict bandwidth densities as one continuous function. The inset in Fig. 4c shows that extrapolating 130 µm (the 2D region) down to 1 µm requires numerous extrapolations, amplifying the error. Extrapolating 65 µm (the 2.xD region) down to 1 µm is more accurate at discrete points compared to extrapolating 130 µm down to 1 µm; however, there is still error in curve fitting at other points along the curve, especially for realizable bandwidth density that can be achieved with correct overhead and power/ground ratios. In summary, we started with two graphs, one for 2D + 2.xD regions down to 65 µm and one for 2.xD + 3D regions from 65 µm down to 1 µm. We tried fitting both graphs in power series, as area scaling to bump pitch is almost a power series, and the corresponding equations appear in the inset of Fig. 4c, one for 65 µm down to 1 µm, representing 2.5D and 3D regions, and the second for the entire region of 130 µm down to 1 µm.

Care needs to be taken in terms of assembly and defects (particle size, clean room) and in future work, it will be possible to reduce the overhead ratio to achieve more bandwidth density. Industry-wide research might always show that the theoretical bandwidth density is sufficiently high, but as intuitive as it may seem that no more bump shrink or new technologies are needed, the need for bandwidth density with practical approaches in real products will help drive the need for further technological improvements and bump shrink with different technologies, and this in turn will also reduce the gap between the theoretical bandwidth density and the realizable bandwidth density at fine pitches.

Data availability

The data that support the plots within this paper and other findings of this study are available from the corresponding author upon reasonable request.

References

1. Nassif, N. et al. (2022) Sapphire rapids: the next-generation Intel Xeon scalable processor. In *2022 IEEE International Solid-State Circuits Conference (ISSCC)* Vol. 65 (ed. Fujino, L. C.) 44–46 (IEEE, 2022); https://doi.org/10.1109/ISSCC42614.2022.9731107
2. Burd, T. et al. Zeppelin: an SoC for multichip architectures. *2019 IEEE J. Solid-State Circuits* **54**, 40–42 (2019).
3. Jiang, H. Intel's Ponte Vecchio GPU: architecture, systems & software. In *2022 IEEE Hot Chips 34 Symposium (HCS)* (ed. Rusu, S.) 1–29 (IEEE, 2022); https://doi.org/10.1109/HCS55958.2022.9895631
4. Choquette, J. NVIDIA Hopper H100 GPU: scaling performance. *IEEE Micro* **43**, 9–17 (2023); https://doi.org/10.1109/MM.2023.3256796
5. Dudeck, G. & Dudeck, J. Design considerations and packaging of a Pentium(R) Pro Processor based multi-chip module for high performance workstation and servers. In *1998 IEEE Symposium on IC/Package Design Integration (Cat. No.98CB36211)* (ed. Sipple, R. S.) 9–15 (IEEE, 2002); https://doi.org/10.1109/IPDI.1998.663613
6. Douglas J. Intel 8×× Series and Paxville Xeon-MP microprocessors. In *2005 IEEE Hot Chips XVII Symposium (HCS)* Vol. 9 (ed. Garb, G.) 1–26 (IEEE, 2005); https://doi.org/10.1109/HOTCHIPS.2005.7476602
7. Mahajan, R. et al., Embedded Multidie Interconnect Bridge—A Localized, High-Density Multichip Packaging Interconnect, in *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 9, 1952-1962, Oct. 2019, https://doi.org/10.1109/TCPMT.2019.2942708
8. Lee, L. C. T. et al. Advanced HDFO packaging solutions for chiplets integration in HPC application. In *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)* (ed. O'Conner, L.) 8–13 (IEEE, 2021); https://doi.org/10.1109/ECTC32696.2021.00013
9. Hou S. Y. et al. Integrated deep trench capacitor in Si interposer for CoWoS heterogeneous integration. In *2019 IEEE International Electron Devices Meeting (IEDM)* (ed. De Salvo, B.) 19.5.1–19.5.4 (IEEE, 2019); https://doi.org/10.1109/IEDM19573.2019.8993498
10. Huang P. K. et al. Wafer level system integration of the fifth generation CoWoS®-S with high performance Si interposer at 2500 mm2. In *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)* (ed. O'Conner, L.) 101–104 (IEEE, 2021); https://doi.org/10.1109/ECTC32696.2021.00028
11. Lau J. H. Recent advances and trends in advanced packaging. *IEEE Trans. Compon.* **12**, 228–252 (2022); https://doi.org/10.1109/TCPMT.2022.3144461
12. Zhang S. et al. Challenges and recent prospectives of 3D heterogeneous integration. *E-Prime-Adv. Electr. Eng. Electron. Energy* **2**, 100052 (2022); https://doi.org/10.1016/j.prime.2022.100052
13. Moore, G. E. Cramming more components onto integrated circuits. *Electronics*, **38**, 114–117 (1965). Reprinted in *IEEE Solid-State Circuits Society Newsletter* **11**, 33–35 (2006); https://doi.org/10.1109/N-SSC.2006.4785860
14. Das Sharma, D. *Universal Chiplet Interconnect Express (UCIe): building an open chiplet ecosystem* White paper (UCIe Consortium, 2022).
15. Universal Chiplet Interconnect Express (UCIe) Specification Rev 1.0, Feb 17, 2022; www.uciexpress.org/specifications
16. Das Sharma, D. System on a package innovations with Universal Chiplet Interconnect Express (UCIe) interconnect. *IEEE Micro* **43**, 76–85 (2023); https://doi.org/10.1109/MM.2023.3235770
17. Das Sharma, D., Pasdast, G., Qian, Z. & Aygun, K. Universal Chiplet Interconnect Express (UCIe): an open industry standard for innovations with chiplets at package level. *IEEE Trans. Compon.* **12**, 1423–1431 (2022); https://doi.org/10.1109/TCPMT.2022.3207195
18. IEEE Electronics Packaging Society Heterogeneous Integration Roadmap, 2021 Edition. *IEEE* (2021). https://eps.ieee.org/hir
19. PCI-SIG, PCI Express® Base Specification revision 5.0, version 1.0, May 28, 2019. https://members.pcisig.com/wg/PCI-SIG/document/13005
20. PCI-SIG, PCI Express® Base Specification revision 6.0, version 1.0, Jan 11, 2022. https://members.pcisig.com/wg/PCI-SIG/document/16609
21. CXL Consortium, Compute Express Link 2.0 Specification. https://members.computeexpresslink.org/wg/CXL/document/1172(2020).
22. Elsherbini, A. et al. Enabling next generation 3D heterogeneous integration architectures on intel process. In *2022 International Electron Devices Meeting (IEDM)* (ed. Hoentschel, J.) 27.3.1–27.3.4 (IEEE, 2022); https://doi.org/10.1109/IEDM45625.2022.10019499
23. Liang, S. W. et al. High performance and energy efficient computing with advanced SoIC™ scaling. In *2022 IEEE 72nd Electronic Components and Technology Conference (ECTC)* (ed. O'Conner, L.) 1090–1094 (IEEE, 2022); https://doi.org/10.1109/ECTC51906.2022.00176
24. Hwang, Y., Moon, S., Nam S. & HoonAhn, J. Chiplet-based system PSI optimization for 2.5D/3D advanced packaging implementation. In *2022 IEEE 72nd Electronic Components and Technology Conference (ECTC)* (ed. O'Conner, L.) 12–17 (IEEE, 2022); https://doi.org/10.1109/ECTC51906.2022.00010
25. Lee, D. U. et al. A 1.2V 8Gb 8-channel 128GB/s high-bandwidth memory (HBM) stacked DRAM with effective microbump I/O test methods using 29nm process and TSV. In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* (ed. Fujino, L. C.) 432–433 (IEEE, 2014); https://doi.org/10.1109/ISSCC.2014.6757501
26. James, D. 3D ICs in the real world. In *25th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC 2014)* (ed. Ne'eman, I.) 113–119 (IEEE, 2014); https://doi.org/10.1109/ASMC.2014.6846988
27. Chiang, C. & Sinha, S. The road to 3D EDA tool readiness. In *2009 Asia and South Pacific Design Automation Conference* (ed. Ishihara, T.) 429–436 (IEEE, 2009); https://doi.org/10.1109/ASPDAC.2009.4796519
28. Natarajan, S. 'Intel Accelerated' Webcast (Replay). *Intel* www.intel.com/content/www/us/en/newsroom/news/intel-accelerated-webcast-livestream-replay.html#gs.ezui39 (2021).
29. Prasad, D. et al. Buried power rails and back-side power grids: Arm® CPU power delivery network design beyond 5nm. In *2019 IEEE International Electron Devices Meeting (IEDM)* (ed. De Salvo, B.) 19.1.1–19.1.4 (IEEE, 2019); https://doi.org/10.1109/IEDM19573.2019.8993617

30. *A case for lowering component level CDM ESD specifications and requirements, Rev 3.0*. White paper (Industry Council on ESD Target Levels, 2021); www.esdindustrycouncil.org/ic/en/shared/industry-council-white-paper-2-rev3.0-may-2021-final.pdf
31. Bailey, B. True 3D-IC Problems. *Semiconductor Engineering* https://semiengineering.com/true-3d-ic-problems/ (2023).

Author contributions

D.D.S. crafted the overall vision and architecture for the next generation UCIe, including its applications and key required characteristics and also did the performance and reliability calculations. G.P. developed the circuit architecture for UCIe-3D and all its corresponding details described in the paper. S.T. performed the bandwidth density, power, performance and reliability calculations and predictions. K.A., with D.D.S., developed the mapping between UCIe-3D architecture and required package technologies and crafted the vision for future SiP architecture using UCIe-3D.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41928-024-01126-y.

Correspondence and requests for materials should be addressed to Debendra Das Sharma.

Peer review information *Nature Electronics* thanks Lihong Cao and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024

 https://doi.org/10.1038/s41928-024-01126-y

Extended Data Table 1 | Key metrics and Characteristics of UCIe 1.0 implementations in standard and advanced packages based on the available technology today

Characteristics / Key Metrics	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 Package (AP)
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). AP max speed: 32G for 45u-55u, 24G for 38u-44u, 16G for 31u-37u, 12G for 25u-30u. Standard has max speed 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)
Bit Error Rate (BER)	< 10^{-27} below 16G for UCIe-a; 12G for UCIe-S; else < 10^{-15}		Probability of a bit error
Reliability (FIT)	0 < FIT (Failure In Time) << 1		FIT: #failures in a billion hours (expecting ~1E-10) w/ UCIe Flit Mode

https://doi.org/10.1038/s41928-024-01126-y

Extended Data Table 2 | Comparison of UCIe-S vs. UCIe-A vs. proposed UCIe-3D. Note that linear bandwidth density is irrelevant in 3D arrangements due to both X and Y degrees of freedom. The 3D bandwidth density includes a very conservative 10% overhead for NoC-level signal repair

UCIe type	Signal integrity features	Termination	Minimum ESD	Power	Max Bandwidth		Latency
			CDM (V)	(pJ/b)	Linear (GB/s/mm)	Aerial (GB/s/mm^2)	(ns)
UCIe-S [14,15]	TX & RX inductor coils, EQ	Resistors for linearity: TX & RX	30	0.5	224	125	2
UCIe-A [14,15]	None	Transistor-only sufficient: TX	30	0.25	1317	1350	2
UCIe-3D*	None	None	0	0.01	N/A[1]	207000[2]	0.5

*Based on our proposed UCIe-3D architecture and analysis provided in this paper

[1]: Linear bandwidth density is irrelevant in 3D arrangements due to both X and Y degrees of freedom

[2]: The 3D bandwidth density includes a very conservative 10% overhead for NoC-level signal repair