

EXHIBIT 803-A-03

Ayşe Coskun et al., *A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems*, ICCAD '18 (2018) ("Coskun")

Ayşe Coskun et al., *A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems*, ICCAD '18 (2018) ("Coskun") anticipates and/or renders obvious (alone or in combination with other references), including at least under Complainant's apparent infringement theory:

- claims 1-3, 7-10 of U.S. Patent No. 11,841,803 ("the '803 patent") under 35 U.S.C. §§ 102 and/or 103.

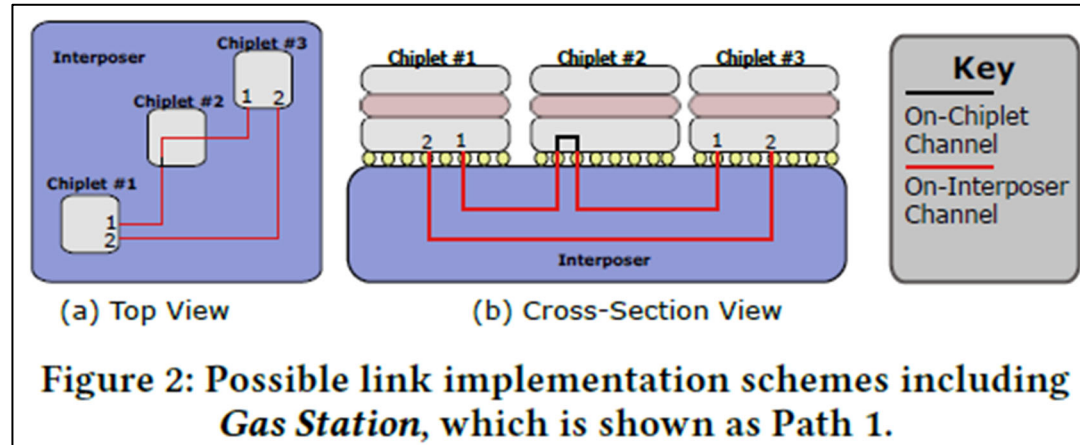
NVIDIA has charted the portions of the reference that correspond to each element, including at least under Complainant's apparent interpretation of the claims as reflected in Complainant's infringement contentions. Where the chart below states that the reference "discloses," "teaches," or "suggests" a claim limitation, such disclosure, teaching, or suggestion may be express, implied, or inherent. The cited evidence is merely illustrative, and NVIDIA reserves the right to cite alternative or additional evidence.

No part of this exemplary chart construes, or is intended to construe, the specification, file history, or claims of the asserted patents. In addition, nothing stated in this chart shall be treated as an admission or suggestion that NVIDIA agrees with Complainant regarding either the scope of any of the asserted claims or Complainant's apparent interpretation of the scope of the asserted claims, or that NVIDIA's accused products meet any limitations of the claims.

To the extent that Complainant contends that Coskun does not disclose one or more limitations of the asserted claims, it would have been obvious to combine the teachings of Coskun with (1) the knowledge of one of ordinary skill in the art, and/or (2) the teachings of any of the prior art set forth in NVIDIA's other invalidity charts and cover pleading, to show all the limitations of the asserted claims, including for the reasons set forth in NVIDIA's Invalidity Contention Cover Pleading, which are incorporated herein.

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
1[Pre]: A system, comprising:	Including at least under Complainant's apparent infringement theory, Coskun discloses and/or renders obvious this limitation. <i>See</i> evidence cited for '803 patent claims 1-3, 7-10. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
1[A]: a central processing unit (CPU) communicably coupled to a first graphics processing unit (GPU) chiplet of a GPU chiplet array,	Including at least under Complainant's apparent infringement theory, Coskun discloses and/or renders obvious this limitation. <i>See, e.g.,</i> Coskun at Fig. 1. Figure 1: The cross-section of a 2.5D system.

Coskun at Fig. 2.



Coskun at Fig. 3.

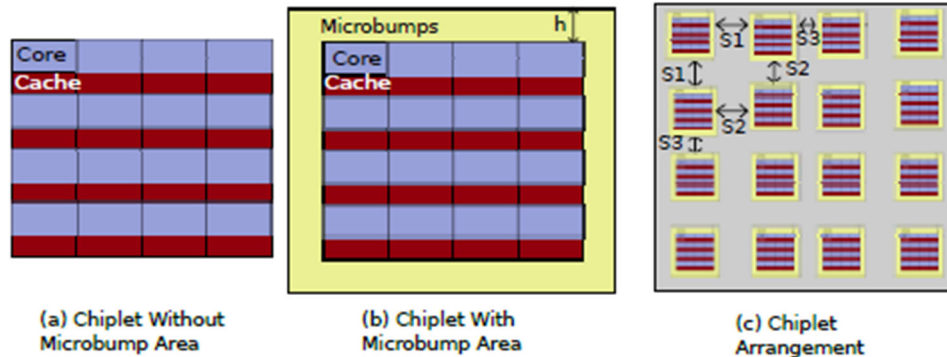


Figure 3: Illustration of the extra microbump area required per chiplet.

Coskun at Fig. 4.

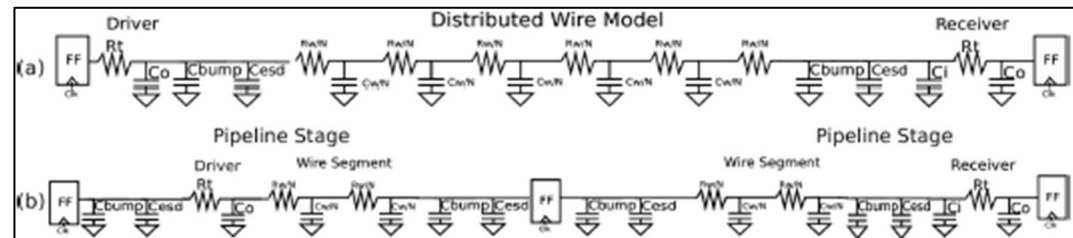
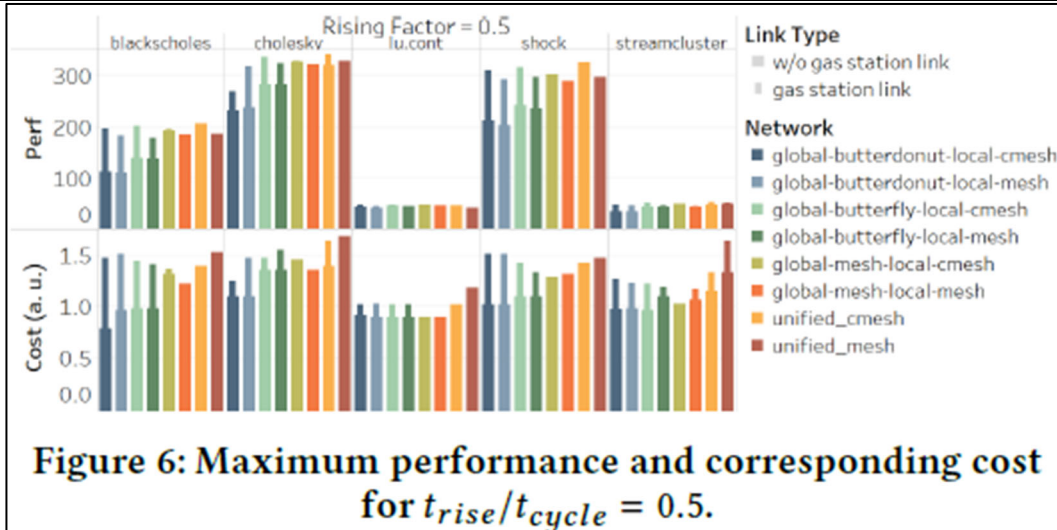
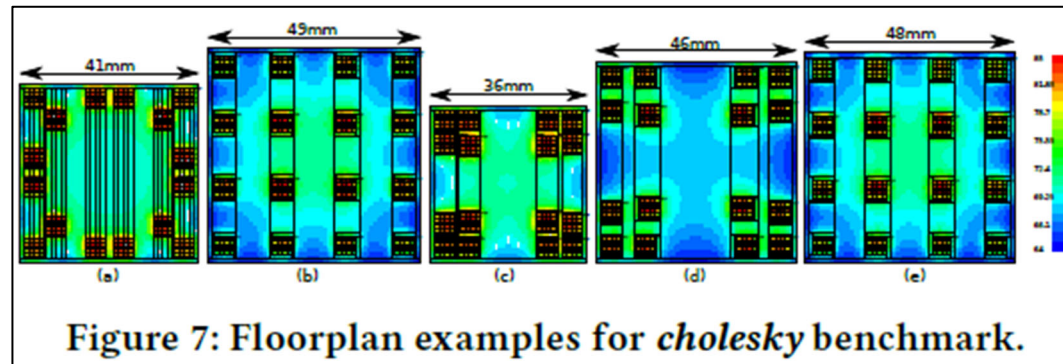


Figure 4: Distributed inter-chiplet link models: (a) repeaterless link and (b) gas station link, in a passive interposer.

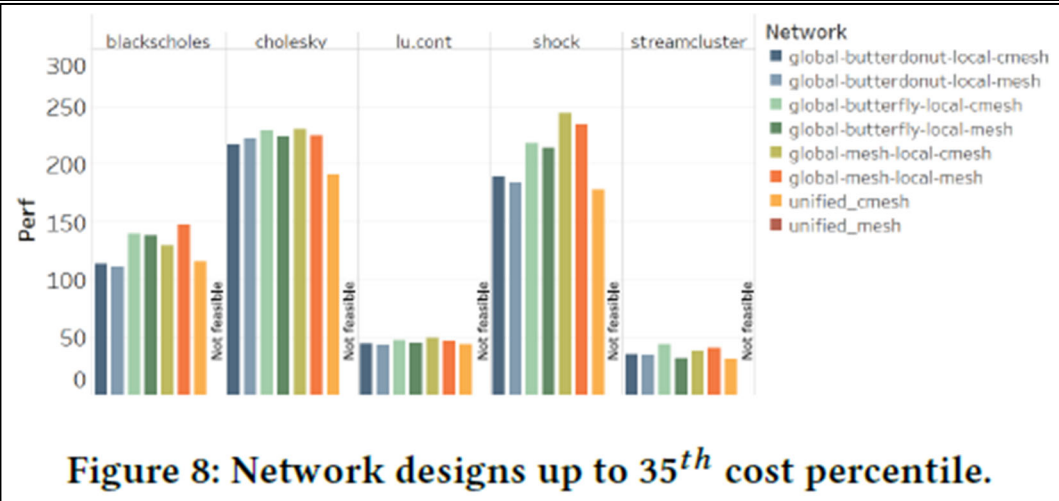
Coskun at Fig. 6.



Coskun Fig. 7.



Coskun Fig. 8.



Coskun at Table 1.

	<i>Unified Mesh</i>	<i>Unified Cmesh</i>	<i>Global Mesh</i>	<i>Global Butterfly</i>	<i>Global Butterdonut</i>	<i>Global Clos</i>
#microbumps	1024	512	256	256	256	2048
h (mm)	0.585	0.315	0.18	0.18	0.18	1.125
Chiplet Size (mm)	5.67	5.13	4.86	4.86	4.86	6.75
Microbump Area Overhead (%)	58.76	29.96	16.64	16.64	16.64	125.0

Table 1: Microbump area overhead for network topologies with shielding overhead included.

Coskun at Table 2.

Technology Node	22nm	65nm
Wire Thickness	300nm	1.5 μ m
Dielectric Height	300nm	0.9 μ m [17]
Wire Width	200nm	1 μ m [26]
C_{bump}	4.5fF	4.5fF [17]
C_{esd}	50fF	50fF [17]
C_{g-t} (Gate Cap)	1.08fF/ μ m	1.05fF/ μ m
C_{d-t} (Drain Cap)	1.5 $\times$ Cg	1.5 $\times$ Cg
R_t (Inverter resistance)	450 $\Omega \cdot \mu$ m	170 $\Omega \cdot \mu$ m
Wire Pitch	0.4 μ m	2 μ m [26]
Flip-Flop Energy per Bit	14fJ/bit [10]	28fJ/bit [18]
Flip-Flop $t_{c-q} + t_{setup}$	49ps [10]	45ps [18]

Table 2: Technology node parameters.

Coskun at Table 3.

(v(V), f(MHz))	$t_{rise}/t_{cycle} = 0.5$			$t_{rise}/t_{cycle} = 0.8$		
	(0.9, 1000)	(0.89, 800)	(0.71, 533)	(0.9, 1000)	(0.89, 800)	(0.71, 533)
1 Cycle	9	11	13	12	15	18
2 Cycles	16	18	23	19	25	30
3 Cycles	21	23	30	25	32	38
4 Cycles	25	27	35	30	38	45
5 Cycles	28	32	41	33	43	52
6+ Cycles	>32	>36	>45	>37	>48	>57

Table 3: Maximum link lengths (in mm) for a given network latency (in cycles), voltage-frequency setting, and rise time constraint.

Coskun at Table 5.

Input	Properties
Chiplets	$ C $ Chiplet instances, at $\{X_c, Y_c\}$ left bottom, $c \in C$. The locations provided for the chiplets are assumed to be legal.
Pin Clumps	$ P $ Pin clump instances of pin capacity P_{th}^{max} each. Each pin clump p has a predetermined location $\{x_p, y_p\}$ relative to the left bottom of the chiplet.
Required Connections	R_{ij} between every pair of chiplets $\{i, j\}$ indicating the number of wires that need to go between the pair of chiplets. If $R_{ij} > 0$ then a net n exists between chiplet i and chiplet j with source $s_n = i$ and sink $t_n = j$.
Routing Rules	Maximum length of a route, D_{max} . Maximum number of segments, S_{max} equal to 1, 2 or 3. $S_{max} \leq 3$ to limit impact on latency.

Table 5: Inputs to the routing optimization.

Coskun at *1 (Abstract) (“2.5D integration technology is gaining popularity in the design of homogeneous and heterogeneous many-core computing systems. 2.5D network design, both inter- and intra-chiplet, impacts overall system performance as well as its manufacturing cost and thermal feasibility. This paper introduces a cross-layer methodology for designing networks in 2.5D systems. We optimize the network design and chiplet placement jointly across logical, physical, and circuit layers to achieve an energy-efficient network, while maximizing system performance, minimizing manufacturing cost, and adhering to thermal constraints. In the logical layer, our cooptimization considers eight different network topologies. In the physical layer, we consider routing, microbump assignment, and microbump pitch constraints to account for the extra costs associated with microbump utilization in the inter-chiplet communication. In the circuit layer, we consider both passive and active links with five different link types, including a gas station link design. Using our cross-layer methodology results in more accurate determination of (superior) inter-chiplet network and 2.5D system designs compared to prior methods. Compared to 2D systems, our approach achieves 29% better performance with the same manufacturing cost, or 25% lower cost with the same performance.”).

Coskun at *1 (§ 1) (“The need to sustain the historical performance and cost scaling in computing systems has led to a growing interest in 2.5D systems [1, 8, 9, 15, 16, 29]. In 2.5D

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	design, multiple chiplets are placed on a silicon interposer, and the chiplets communicate using links integrated into the interposer. 2.5D integration technology provides multiple potential benefits compared to 2D systems, including greater system performance within thermal constraints [12], heterogeneous integration of multiple technologies [1, 6], and reduced overall system cost [16]. However, 2.5D integration technology also opens up a number of design challenges, ranging from circuit and physical challenges (design and routing of inter-chiplet links, placement and floorplanning of chiplets on the interposer, microbump assignment, etc.) to architectural and system-level challenges (design of the inter-chiplet network architecture, partitioning a system into 2.5D-integrated heterogeneous functional components, etc.). Many other technical and business challenges, including design for thermomechanical stress, test strategy, and supply chain structure, are identified by Radojcic [26].”). Coskun at *1 (§ 1) (“In this paper, we perform a cross-layer co-optimization of 2.5D inter-chiplet network design and chiplet placement. Our cooptimization methodology focuses on network topologies, link circuit options, and microbump pitch- and interconnect RC-aware routing of links. It maximizes performance and/or minimizes cost at the system level, while satisfying system power and thermal constraints. The need for such a cross-layer methodology as ours can be easily seen by considering the following. If we adopt a topdown approach, an architecture analysis of network topologies tells us that high-radix, low-diameter networks should be used for inter-chiplet networks, as they provide the best overall system performance (in instructions per cycle). However, in the physical layer, realization of high-radix, low-diameter networks requires long wires, which can limit the network performance and, hence, the overall system performance. Using repeaters on long wires to improve performance would necessitate active (rather than passive) interposer technology. Since active interposers are 10× more expensive than passive interposers [25], the system cost equation changes and the top-down intuition is flawed. On the other hand, a bottom-up, cost-centric perspective prompts the use of a passive interposer, which can only support repeaterless links in the circuit layer, thus limiting link performance and maximum link length. This leads to the adoption of low-radix, high-diameter networks at the inter-chiplet level, which lowers overall system performance.”).

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	Coskun at *1-*2 (§ 1) (“Our work fills a significant gap in the literature on inter-chiplet network design and floorplan/placement optimization of 2.5D systems. No prior work has simultaneously considered thermal behavior of chiplets, multiple potential network topologies, multiple inter-chiplet link options, and physical design constraints associated with routing these links. Thus, previous approaches can incorrectly evaluate cost, performance, power and thermal feasibility, as well as other important parameters of 2.5D system solutions. Consequently, there is a risk of identifying suboptimal inter-chiplet network and 2.5D system floorplan solutions, which can lead to inefficient architectural decisions. For example, in our recent work [12], we describe a methodology to place chiplets (connected by a mesh) that results in thermally-safe, high-performance, and lowcost 2.5D systems. However, in the logical layer, we only consider a Unified-Mesh1 network topology. In the physical layer, there is no accounting of the area overhead associated with the microbumps required for the inter-chiplet links. In the circuit layer, we consider only one type of link. As elaborated below, the present paper shows that a careful accounting of microbump overhead, along with consideration of multiple network topology options and link design options, leads to a solution that can achieve 16% higher performance at comparable cost, and/or 18% lower cost at comparable performance, with respect to our prior best solutions.”). Coskun at *2 (§ 1) (“The main contributions of this paper are as follows: • We develop a cross-layer co-optimization methodology that optimizes inter-chiplet network design jointly with chiplet placement across logical, physical, and circuit layers. Our methodology optimizes a given 2.5D system for performance, cost, and wirelength, while ensuring that it is thermally safe. The outcome of the co-optimization comprises placement of chiplets on the interposer, logical topology of the interchiplet network, and circuit design and routing of the links that form the network. • Our co-optimization considers a rich solution space. (i) In the logical layer, we consider a variety of Global, Local, and Unified network topologies. (ii) In the physical layer, we incorporate well-calibrated microbump overhead models into our area and cost models. We further consider the finite density of microbumps per unit die area,

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	and assess achievable physical wiring distances (hence, achievable link latencies). (iii) In the circuit layer, we explore repeaterless non-pipelined, repeaterless pipelined, repeatered pipelined, and repeatered non-pipelined types of inter-chiplet links. We further consider a gas station link design to enable pipelining in passive inter-chiplet links. • Our heuristic-based cross-layer co-optimization has several novel elements. (i) For a given chiplet placement and network topology, we perform routing and microbump assignment using a flow-based mixed integer-linear program (MILP) to minimize the maximum link latency. (ii) We use workload and network throughput-aware thermal simulation outputs from HotSpot [32] to assess the thermal feasibility of placement and network topology solutions. (iii) We apply simulated annealing to search over our high-dimensional system solution space.”). Coskun at *2 (§ 2) (“In the logical layer, Kannan et al. [16] have evaluated various logical topologies for 2.5D systems, but their work does not consider microbump area overheads, different inter-chiplet link options, or physical implementation of the 2.5D layout. Ahmed et al. [2] propose a hierarchical mesh network for inter-chiplet communication. Both Kannan et al. [16] and Ahmed et al. [2] assume a “minimally active” interposer, which could be unrealistic from a cost perspective (see Section 3). Akgun et al. [3] evaluate three specialized memory-to-core network topologies, yet the evaluation is limited to the logical layer with a static placement of chiplets, and implications of design choices on physical and circuit layers are not explored. None of these works take thermal effects into account or perform a physical design optimization of the 2.5D inter-chiplet network.”). Coskun at *2 (§ 2) (“In the physical layer, Funke et al. [14] have proposed various algorithms that exhaustively search for optimal placement and routing solutions for up to six chiplets. The recent work of Osmonolovskyi et al. [24] handles up to 11-chiplet design complexity using pruning methods. Minz et al. [22] and Fang et al. [13] focus on routing of inter-chiplet links on an interposer. Liu et al. [20] aim to reduce the number of metal layers in the interposer. These works do not consider thermal effects while finding placement and routing solutions. In our

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	prior work [12], we propose a thermally-aware chiplet placement solution. However, our prior work does not perform routing and only computes a placement solution.”). Coskun at *2 (§ 2) (“In contrast to these previous works, our methodology jointly considers logical, physical, and circuit design of the inter-chiplet network. We evaluate a variety of logical topologies, while being aware of the network design feasibility in both the circuit layer and the physical layer. In the circuit layer, we evaluate various link design options. In the physical layer, we develop a thermally-aware placement and routing solution. Our cross-layer methodology, thus, obtains 2.5D system solutions that, having more complete and accurate modeling foundations, come closer to defining the true envelope of 2.5D system performance and cost under power and thermal constraints.”). Coskun at *2 (§ 3) (“A cross-layer inter-chiplet network design methodology must comprehend a vast design space that spans the logical, physical and circuit design layers. In this section, we describe the design space for each of these three layers, along with key parameters of interest.”). Coskun at *2-*3 (§ 3.1) (“Each core, together with its L1 and L2 caches, has a square layout. Following our priorwork [12],we study chiplet-based integration of 16 identical chiplets on an interposer, where each chiplet contains 16 cores. Figure 1 illustrates the cross-section view of a 2.5D system. We assume that the 22nm chiplets are placed on an interposer that is designed in 65nm process technology. Microbumps connect the chiplets to the interposer substrate. The system is placed on a System-in-Package (SiP) substrate, with C4 (“flip chip bumps”) connecting the interposer to the SiP substrate. We enable direct comparison of this work with our prior work [12] by designing a Unified-Mesh network using our cross-layer methodology. It should be noted that our broader conclusions are agnostic of the specific core count, core architecture, and technology nodes for the chiplets and the interposer.”). Coskun at *3 (§ 3.2) (“In the logical layer, we explore several different network topologies [35]. We limit the intra-chiplet network to Local-Mesh and Local-Cmesh topologies. For the

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	inter-chiplet network, we design and evaluate Global-Butterfly, Global-Butterdonut [16], and Global-Mesh topologies. For the Unified networks, we evaluate Unified-Mesh and Unified-Cmesh.”). Coskun at *3 (§ 3.3) (“Physical design of the inter-chiplet network consists of placement of the chiplets, along with a routing solution connecting the chiplets that is consistent with the chosen network topology (see Section 3.2). The placement of chiplets affects the temperature map and the length of the links among chiplets, while the routing solution in turn affects the microbump assignment and circuit choices for the link. Further, we explicitly account for the area overhead of microbumps and the associated inter-chiplet drivers and receivers placed along peripheral regions of the chiplets.” (footnote omitted)). Coskun at *3 (§ 3.3) (“Inter-chiplet links can be routed on a passive or an active interposer. Microbumps and ESD protection are required at the beginning and the end of links that go through interposers, and this design constraint adds capacitance [17]. Passive interposers cost less due to their lower manufacturing cost and higher yield [25]. Active interposers allow for repeaters and/or flip-flops (for pipelining) on the interposer. This enables better link bandwidth and latency at the expense of higher manufacturing cost [25]. We conduct a preliminary study of the performance benefit of an active interposer. We observe 2× to 3× latency improvements for the same link length and 50% longer links for the same throughput, but this comes at a 10× cost overhead (\$500 per wafer for passive interposer vs. \$5000 per wafer for active interposer [25]). Given this cost overhead, we rule out active interposers as a realistic option in the near term, and do not consider this option in our present study.”). Coskun at *3 (§ 3.3) (“Passive interposers limit the bandwidth of the signal by degrading rise/fall times. Hence, we use a gas station link, where we can “refuel” a passive link using repeaters and/or flip-flops that are inside other chiplets along the way from the source chiplet to the sink chiplet. Figure 2 shows two implementation schemes for a chiplet-to-chiplet link. Figure 2(a) shows the top view of the paths for the two links connecting Chiplet #1 to Chiplet #3, which are far (e.g., > 10mm) from each other. Figure 2(b) shows a cross-sectional view of

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	the two paths between Chiplet #1 and Chiplet #3. Path 1 uses Chiplet #2 as a gas station, while Path 2 is a direct connection without any gas station. It is important to note the differences between an inter-chiplet repeaterless pipelined link and a gas station link. (i) Pipelining repeaterless links requires an active interposer, while for gas station links we can use a passive interposer. (ii) Active elements required for repeaterless pipelined links are designed using the active interposer's technology node, while active elements required for gas station links are designed using the chiplet's technology node. (iii) Using gas station links requires additional microbumps, and in turn, has an area overhead.”). Coskun at *3 (§ 3.3) (“When considering 2.5D inter-chiplet links, recent works have overlooked the microbump overhead while assessing 2.5D integration benefits. Generally, the number of required microbumps will change according to the network topology. An increase in the number of inter-chiplet links increases the number of required microbumps. Further, additional microbumps (20% according to Radojcic [26]) must be reserved for power delivery and signal shielding purposes. Figure 3 shows the chiplet without and with the extra area required for microbumps. Table 1 presents the overhead due to microbumps for different network topologies designed using repeaterless non-pipelined links. The calculations are for the 256-core system divided across 16 chiplets, with each chiplet having an area of 4.5mm × 4.5mm, and a microbump pitch of 45um. Here, h indicates the width of the extra space along the chiplet periphery required for the microbumps used for the inter-chiplet links [26]. The use of gas station link design will further increase microbump count. We do not list the microbump area overhead associated with use of gas station links since this depends upon the placement solution as well as the network type.”). Coskun at *3 (§ 3.4) (“There are multiple circuit design options for inter-chiplet links. For passive interposers, the link on the interposer itself is repeaterless, but with the inclusion of gas stations, the link can use repeaters and/or flip-flops (for pipelining) in intermediate chiplets to regenerate and retime the signal. We limit t_{rise}/t_{cycle} to less than 0.5, to ensure full voltage swing at all nodes in the presence of non-idealities such as supply noise and jitter. We also explore t_{rise}/t_{cycle} of 0.8 that allows us to go longer distances without repeaters. Relaxing the

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	clock period or allowing for multi-cycle bit-periods permits us to use longer inter-chiplet links.”). Coskun at *4 (§ 3.4) (“Figure 4 shows distributed circuit models for link types; (a) repeaterless link in passive interposer, and (b) gas station link in passive interposer. We model wire parasitics using a distributed, multi-segment π model. We use 22nm technology parameters for intra-chiplet components (drivers, receivers, repeaters, and flipflops of the links), while we use 65nm parameters for the interchiplet components of the links. Table 2 shows technology parameter values used in our experiments. We calculate capacitance and resistance based on the model in Wong et al. [30], and we calibrate our stage and path delay estimates based on extraction from layout and Synopsys PrimeTime timing reports.”). Coskun at *4 (§ 4) (“In this section, we describe how we optimize the network design across the layers described in Section 3, using a cross-layer approach. We show our evaluation framework in Figure 5. We first construct oracles for system performance, cost, and interconnect performance. Each of these oracles gives us an element (performance, cost, and latency) of the co-optimization function. Our method for finding a placement solution of chiplets uses a simulated annealing algorithm. We build a search and sort engine that places the oracles and the placement algorithm in a loop to search for a solution across the logical, physical and circuit layers. Table 4 shows the notations we use in the various steps of our crosslayer co-optimization methodology. The placement algorithm uses HotSpot to determine the thermal profile and an MILP is used to find the optimal routing solution. Thus, we determine the feasibility of each placement using HotSpot simulations and the MILP solution.”). Coskun at *4 (§ 4.1) (“We build a system performance oracle that tells us the overall system performance and total core power for a given network topology, voltage-frequency setting, and link latency. To create the oracle, we use Sniper [7] to precompute system performance for a variety of network topologies, voltage-frequency settings, and link latencies. Our system architecture is the 256-core architecture described in Section 3.1. Eight memory controllers are

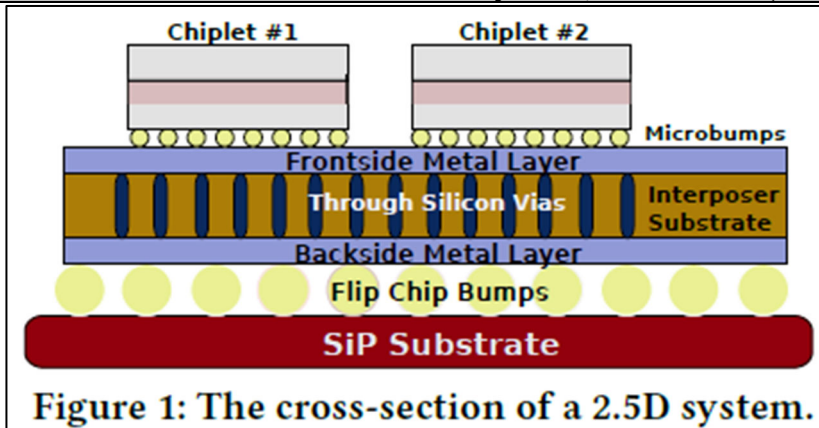
U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	placed next to the top and bottom rows of cores. We implement the inter-chiplet and intra-chiplet network models discussed in Section 3.2 using either passive links or gas station links (see Section 3.3). For passive links without gas stations, we vary inter-chiplet latency values from 1 to 5 cycles, and for gas station links we consider 2- or 3-stage pipelined links. We apply three voltage-frequency settings, (0.9V, 1000MHz), (0.89V, 800MHz) and (0.71V, 533MHz). We fast-forward sequential initialization regions and simulate up to 10 billion instructions in the region of interest using Sniper, with all 256 cores active, to collect performance statistics for five benchmarks. This takes 1.7k CPU hours. We use McPAT [19] to convert the performance results to power traces needed for generating the thermal profile.”). Coskun at *4 (§ 4.2) (“We build a cost oracle that tells us the manufacturing cost of 2.5D systems for a given interposer size, network topology and gas station stage count. We adopt the 2.5D cost model proposed by Stow et al. [28], which takes the cost and yield of chiplets, interposer, and microbump bonding into account, assuming known good dies. We compute the cost of various interposer sizes from 20mm × 20mm to 50mm × 50mm. To estimate the chiplet cost, we compute the number of microbumps required for different network topologies and gas station stages, determine the corresponding chiplet area overhead, and from all these we then calculate the manufacturing cost.” (footnote omitted)). Coskun at *5 (§ 4.4) (“We use simulated annealing to find a placement that meets the thermal constraint and the maximum link length constraint as evaluated by HotSpot and the routing MILP (with the maximum values provided by the interconnect performance oracle), respectively. We assume a symmetric layout similar to that used in our prior work [12]. As shown in Figure 3(c), we use {s1, s2, s3} as the spacings between chiplets. Simulated annealing searches the solution space in the manner shown in Figure 5. The placement optimization also estimates the microbump area overhead based on the routing solution, link type, and network choice.”). Coskun at *5 (§ 4.4.2) (“We build an MILP that takes the placement of chiplets and the logical network connections as input, and provides the optimal routing solution, including microbump

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	assignment, as an output. The routing optimization is performed internally in the placement optimization as seen in Figure 5. The objective of the MILP is a weighted function of the maximum length of a route on the interposer and the total routing area overhead. We frame the delivery of required numbers of wires between chiplets as multi-commodity flow, and formulate an MILP to find optimal routing solutions that comprehend the finite availability of microbumps in regions of the chiplet periphery.”). Coskun at *5-*6 (§ 4.4.2) (“Based on the inputs to the routing optimization step (see Table 5), we precompute d_{ihjk}, the routing distance (assuming Manhattan routing) from pin clump h on chiplet i to pin clump k on chiplet j, using Equation (1). Equation (2) is the objective function for the MILP that includes the maximum length L, and the total length of the routes. In all reported experiments, we set $\alpha = 1$ and $\beta = 0$. Equation (3) imposes an upper bound on L, ensuring that the solution has routes satisfying the input maximum-length constraint $D_{\max}$. Equation (4) ensures that the flow variable $f^{n_{ihjk}}$ is a nonnegative number. Equation (5) is the flow constraint governing the flow variables $f^{n_{ihjk}}$. It ensures the sum of all flows for a net n, over all pin clumps from chiplet s_n to chiplet t_n, meets the R_{ij} requirement. It also ensures that net flow is 0 for all other (nonsource, non-sink) chiplets for the given net. Equation (6) ensures that there is no input flow (for net n) for the source chiplet of net n. Similarly, Equation (7) ensures that there is no output flow from the sink chiplet of net n. Equation (8) ensures that the sum of input and output flows from a given pin clump is always less than or equal to the capacity of the pin clump. This ensures that all routes have available pins. Equation (9) defines $\lambda^{n_{ihjk}}$ as a boolean value based on $f^{n_{ihjk}}$. This helps identify the maximum route length L, as shown in Equation (10). Equation (11) constrains the maximum number of segments ($S_{\max}$) to be either 1, 2 or 3. If $S_{\max} = 1$, no gas stations are permitted, while if $S_{\max} = 2$ or $S_{\max} = 3$, then gas stations are permitted, allowing for 1 or 2 gas station hops, respectively.”). Coskun at *7 (§ 5) (“In this section, we discuss the results of application of our proposed cross-layer co-optimization methodology. We run multithreaded workloads from SPLASH-2 (cholesky, lu.cont) [31], PARSEC (blackscholes, streamcluster) [4], and UHPC (shock) [5] to get a variety of power and performance profiles. For each benchmark, we determine the

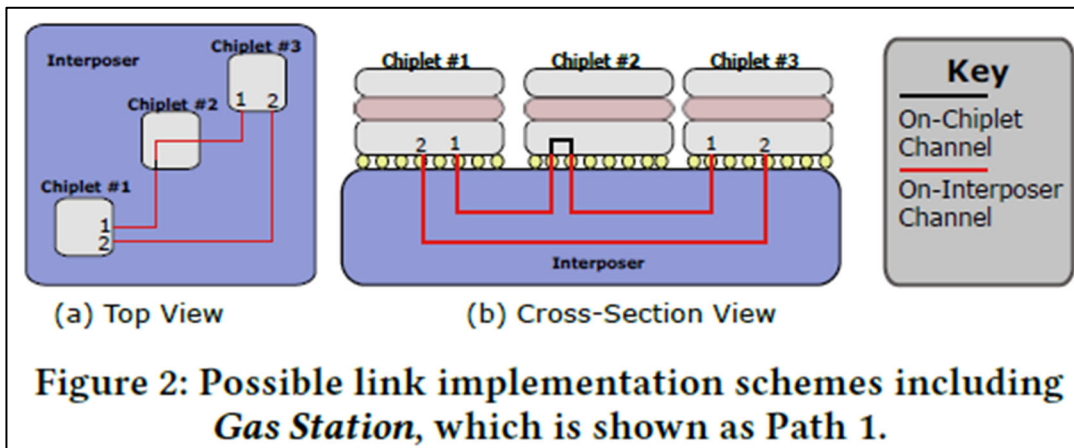
U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	chiplet placement solution, network routing solution, link type, voltage-frequency setting and network topology. In Figure 6, we show the maximum achievable performance and the corresponding cost of all networks across the five benchmarks for $t_{rise}/t_{cycle} = 0.5$. We show results with and without gas stations.”). Coskun at *7 (§ 5) (“If we do not use gas station links, Unified-Mesh outperforms other networks when running cholesky and streamcluster by 1% to 39%. Unified-Cmesh outperforms all other networks for the remaining benchmarks by <1% to 85%. The higher performance of Unified-Mesh/Cmesh is because they have shorter inter-chiplet links and so they easily achieve single-cycle latency even without gas stations. The latency penalty of long links in Global-Butterfly-Local-Mesh/Cmesh and Global-Butterdonut-Local-Mesh/Cmesh leads to lower performance. On average Unified-Cmesh network has the best performance among all networks. It has more inter-chiplet channels compared to Global networks that results in less contention in the inter-chiplet links, and at the same time it has lower hop count than Unified-Mesh that results in lower latency. The higher performance of Unified-Mesh/Cmesh comes at a cost. Unified-Mesh network is the most expensive and has a manufacturing cost that is 6% to 90% higher than other networks.”). Coskun at *7 (§ 5) (“To better understand the design space, we also evaluate maximum performance and corresponding cost for networks with and without gas station links when t_{rise}/t_{cycle} is 0.8. With this t_{rise}/t_{cycle}, longer inter-chiplet link lengths without gas stations are feasible. The relaxed length constraint also reduces the microbump and pipeline stage count, which reduces the cost. For t_{rise}/t_{cycle} of 0.8, without gas stations, Unified-Cmesh outperforms other networks by <1% to 47%. Unified-Mesh has the highest cost and it is 4% to 52% greater than that of other networks. With gas stations, the performance of Unified-Cmesh is <1% to 11% greater than other networks. Unified-Mesh has the highest cost for all benchmarks except blackscholes and shock, and it is 8% to 60% higher than the cost of the remaining networks. For blackscholes, Global-Butterfly-Local-Cmesh has the highest cost. Here the cost is 2% to

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	18% higher than the remaining networks. For shock, Global-Butterdonut-Local-Cmesh has the highest cost and it is 2% to 20% higher than the remaining networks.”). Coskun at *8 (§ 5) (“Finally, we discuss the power of the inter-chiplet network. We see that the highest inter-chiplet network utilization is seen when we run shock on a Unified-Cmesh network. While running shock on Unified-Cmesh, inter-chiplet network power is at most 2% of the overall system power. Theoretically, in very highly threaded applications of the future, we could get much higher network utilizations and then the power of the inter-chiplet network would become a concern.” (footnote omitted)). Coskun at *8 (§ 6) (“In this paper, we have introduced a cross-layer co-optimization methodology for inter-chiplet network design and chiplet placement in 2.5D systems. We have jointly considered network design in the logical, physical, and circuit layers to determine the optimal network choices, link choices, chiplet placements, and link routes to achieve a multi-objective co-optimization goal. We have also proposed to use a gas station link design to enable pipelined interchiplet links when using a passive cost-effective interposer. Our optimization has leveraged well-calibrated models of prior work. We have demonstrated that, compared to 2D systems, our optimized 2.5D systems can achieve 29% better performance with the same manufacturing cost, or 25% lower cost with the same performance.”). Coskun at *8 (§ 6) (“Throughout this work, we have focused on running a single parallel application at a time and have shown the co-optimization outcomes for a variety of benchmarks. Based on these results, a 2.5D system can be further optimized in an application-aware manner (e.g., based on specific applications or worst-/average-case results). Interesting open problems include co-optimization with multi-application scenarios, allocation of threads in a network-aware manner, co-optimization with heterogeneous chiplets, and exploration of active interposer. Also, while we have designed our system for the worst-case link latencies under a global latency constraint, future work involves designing networks with variable link latencies.”).

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
1[B]: wherein the GPU chiplet array includes: the first GPU chiplet communicably coupled to the CPU via a bus; and	Including at least under Complainant's apparent infringement theory, Coskun discloses and/or renders obvious this limitation. <i>See</i> evidence cited for '803 patent limitation 1[A]. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
1[C]: a second GPU chiplet communicably coupled to the first GPU chiplet via a passive crosslink,	Including at least under Complainant's apparent infringement theory, Coskun discloses and/or renders obvious this limitation. <i>See, e.g.,</i> Coskun at Fig. 1.



Coskun at Fig. 2.



Coskun at Fig. 3.

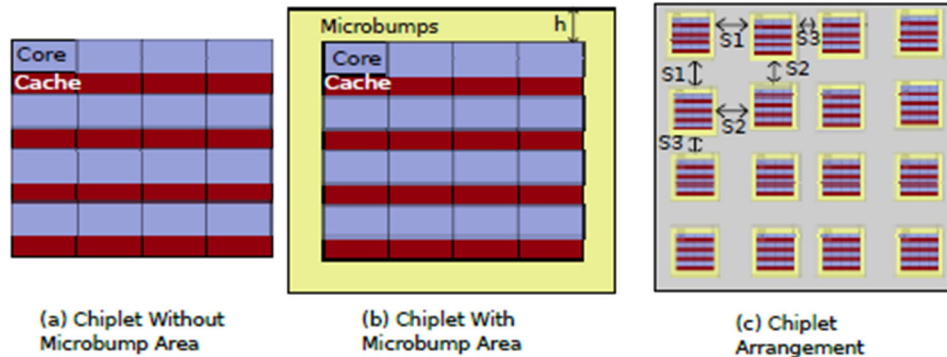


Figure 3: Illustration of the extra microbump area required per chiplet.

Coskun at Fig. 4.

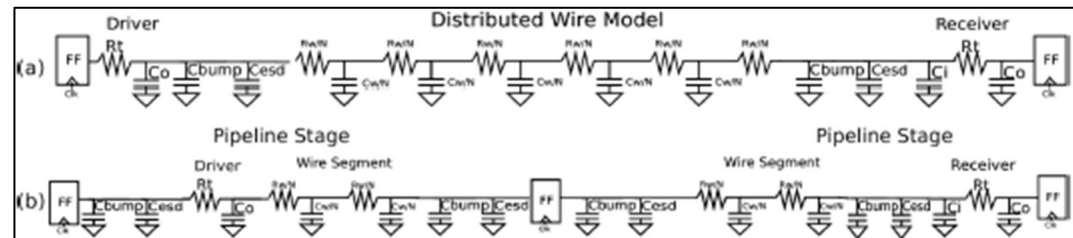
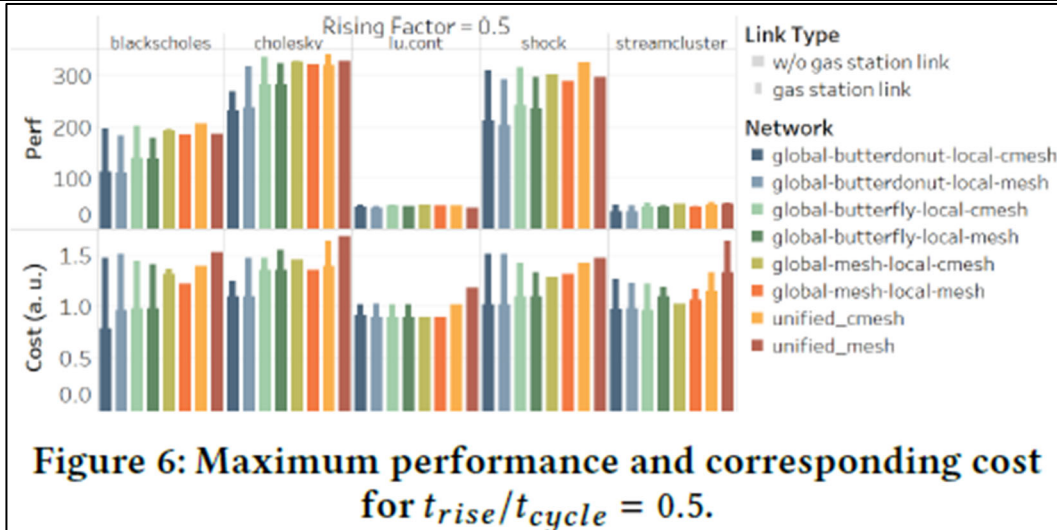
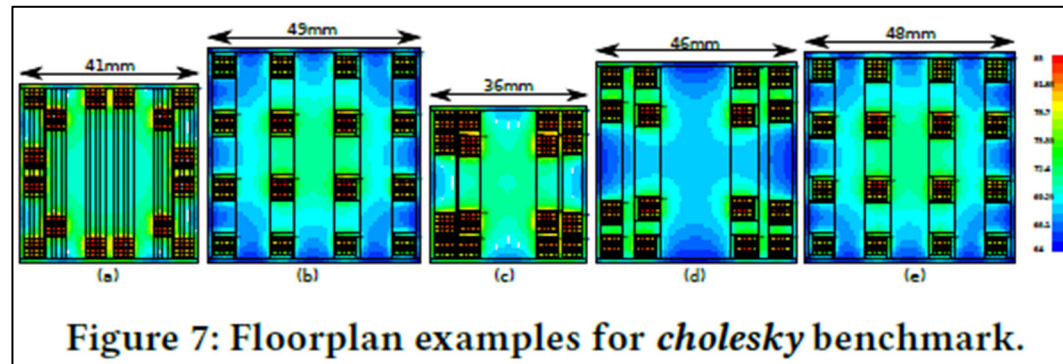


Figure 4: Distributed inter-chiplet link models: (a) repeaterless link and (b) gas station link, in a passive interposer.

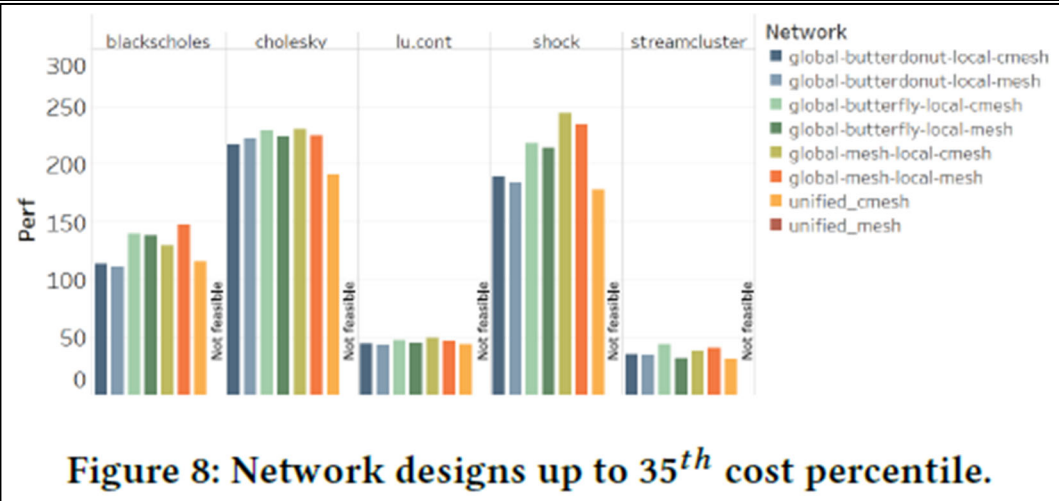
Coskun at Fig. 6.



Coskun Fig. 7.



Coskun Fig. 8.



Coskun at Table 1.

	<i>Unified Mesh</i>	<i>Unified Cmesh</i>	<i>Global Mesh</i>	<i>Global Butterfly</i>	<i>Global Butterdonut</i>	<i>Global Clos</i>
#microbumps	1024	512	256	256	256	2048
h (mm)	0.585	0.315	0.18	0.18	0.18	1.125
Chiplet Size (mm)	5.67	5.13	4.86	4.86	4.86	6.75
Microbump Area Overhead (%)	58.76	29.96	16.64	16.64	16.64	125.0

Table 1: Microbump area overhead for network topologies with shielding overhead included.

Coskun at Table 2.

Technology Node	22nm	65nm
Wire Thickness	300nm	1.5 μ m
Dielectric Height	300nm	0.9 μ m [17]
Wire Width	200nm	1 μ m [26]
C_{bump}	4.5fF	4.5fF [17]
C_{esd}	50fF	50fF [17]
C_{g-t} (Gate Cap)	1.08fF/ μ m	1.05fF/ μ m
C_{d-t} (Drain Cap)	1.5 $\times$ Cg	1.5 $\times$ Cg
R_t (Inverter resistance)	450 $\Omega \cdot \mu$ m	170 $\Omega \cdot \mu$ m
Wire Pitch	0.4 μ m	2 μ m [26]
Flip-Flop Energy per Bit	14fJ/bit [10]	28fJ/bit [18]
Flip-Flop $t_{c-q} + t_{setup}$	49ps [10]	45ps [18]

Table 2: Technology node parameters.

Coskun at Table 3.

(v(V), f(MHz))	$t_{rise}/t_{cycle} = 0.5$			$t_{rise}/t_{cycle} = 0.8$		
	(0.9, 1000)	(0.89, 800)	(0.71, 533)	(0.9, 1000)	(0.89, 800)	(0.71, 533)
1 Cycle	9	11	13	12	15	18
2 Cycles	16	18	23	19	25	30
3 Cycles	21	23	30	25	32	38
4 Cycles	25	27	35	30	38	45
5 Cycles	28	32	41	33	43	52
6+ Cycles	>32	>36	>45	>37	>48	>57

Table 3: Maximum link lengths (in mm) for a given network latency (in cycles), voltage-frequency setting, and rise time constraint.

Coskun at Table 5.

Input	Properties
Chiplets	$ C $ Chiplet instances, at $\{X_c, Y_c\}$ left bottom, $c \in C$. The locations provided for the chiplets are assumed to be legal.
Pin Clumps	$ P $ Pin clump instances of pin capacity P_{th}^{max} each. Each pin clump p has a predetermined location $\{x_p, y_p\}$ relative to the left bottom of the chiplet.
Required Connections	R_{ij} between every pair of chiplets $\{i, j\}$ indicating the number of wires that need to go between the pair of chiplets. If $R_{ij} > 0$ then a net n exists between chiplet i and chiplet j with source $s_n = i$ and sink $t_n = j$.
Routing Rules	Maximum length of a route, D_{max} . Maximum number of segments, S_{max} equal to 1, 2 or 3. $S_{max} \leq 3$ to limit impact on latency.

Table 5: Inputs to the routing optimization.

Coskun at *1 (Abstract) (“2.5D integration technology is gaining popularity in the design of homogeneous and heterogeneous many-core computing systems. 2.5D network design, both inter- and intra-chiplet, impacts overall system performance as well as its manufacturing cost and thermal feasibility. This paper introduces a cross-layer methodology for designing networks in 2.5D systems. We optimize the network design and chiplet placement jointly across logical, physical, and circuit layers to achieve an energy-efficient network, while maximizing system performance, minimizing manufacturing cost, and adhering to thermal constraints. In the logical layer, our cooptimization considers eight different network topologies. In the physical layer, we consider routing, microbump assignment, and microbump pitch constraints to account for the extra costs associated with microbump utilization in the inter-chiplet communication. In the circuit layer, we consider both passive and active links with five different link types, including a gas station link design. Using our cross-layer methodology results in more accurate determination of (superior) inter-chiplet network and 2.5D system designs compared to prior methods. Compared to 2D systems, our approach achieves 29% better performance with the same manufacturing cost, or 25% lower cost with the same performance.”).

Coskun at *1 (§ 1) (“The need to sustain the historical performance and cost scaling in computing systems has led to a growing interest in 2.5D systems [1, 8, 9, 15, 16, 29]. In 2.5D

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	design, multiple chiplets are placed on a silicon interposer, and the chiplets communicate using links integrated into the interposer. 2.5D integration technology provides multiple potential benefits compared to 2D systems, including greater system performance within thermal constraints [12], heterogeneous integration of multiple technologies [1, 6], and reduced overall system cost [16]. However, 2.5D integration technology also opens up a number of design challenges, ranging from circuit and physical challenges (design and routing of inter-chiplet links, placement and floorplanning of chiplets on the interposer, microbump assignment, etc.) to architectural and system-level challenges (design of the inter-chiplet network architecture, partitioning a system into 2.5D-integrated heterogeneous functional components, etc.). Many other technical and business challenges, including design for thermomechanical stress, test strategy, and supply chain structure, are identified by Radojcic [26].”). Coskun at *1 (§ 1) (“In this paper, we perform a cross-layer co-optimization of 2.5D inter-chiplet network design and chiplet placement. Our cooptimization methodology focuses on network topologies, link circuit options, and microbump pitch- and interconnect RC-aware routing of links. It maximizes performance and/or minimizes cost at the system level, while satisfying system power and thermal constraints. The need for such a cross-layer methodology as ours can be easily seen by considering the following. If we adopt a topdown approach, an architecture analysis of network topologies tells us that high-radix, low-diameter networks should be used for inter-chiplet networks, as they provide the best overall system performance (in instructions per cycle). However, in the physical layer, realization of high-radix, low-diameter networks requires long wires, which can limit the network performance and, hence, the overall system performance. Using repeaters on long wires to improve performance would necessitate active (rather than passive) interposer technology. Since active interposers are 10× more expensive than passive interposers [25], the system cost equation changes and the top-down intuition is flawed. On the other hand, a bottom-up, cost-centric perspective prompts the use of a passive interposer, which can only support repeaterless links in the circuit layer, thus limiting link performance and maximum link length. This leads to the adoption of low-radix, high-diameter networks at the inter-chiplet level, which lowers overall system performance.”).

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	Coskun at *1-*2 (§ 1) (“Our work fills a significant gap in the literature on inter-chiplet network design and floorplan/placement optimization of 2.5D systems. No prior work has simultaneously considered thermal behavior of chiplets, multiple potential network topologies, multiple inter-chiplet link options, and physical design constraints associated with routing these links. Thus, previous approaches can incorrectly evaluate cost, performance, power and thermal feasibility, as well as other important parameters of 2.5D system solutions. Consequently, there is a risk of identifying suboptimal inter-chiplet network and 2.5D system floorplan solutions, which can lead to inefficient architectural decisions. For example, in our recent work [12], we describe a methodology to place chiplets (connected by a mesh) that results in thermally-safe, high-performance, and lowcost 2.5D systems. However, in the logical layer, we only consider a Unified-Mesh1 network topology. In the physical layer, there is no accounting of the area overhead associated with the microbumps required for the inter-chiplet links. In the circuit layer, we consider only one type of link. As elaborated below, the present paper shows that a careful accounting of microbump overhead, along with consideration of multiple network topology options and link design options, leads to a solution that can achieve 16% higher performance at comparable cost, and/or 18% lower cost at comparable performance, with respect to our prior best solutions.”). Coskun at *2 (§ 1) (“The main contributions of this paper are as follows: • We develop a cross-layer co-optimization methodology that optimizes inter-chiplet network design jointly with chiplet placement across logical, physical, and circuit layers. Our methodology optimizes a given 2.5D system for performance, cost, and wirelength, while ensuring that it is thermally safe. The outcome of the co-optimization comprises placement of chiplets on the interposer, logical topology of the interchiplet network, and circuit design and routing of the links that form the network. • Our co-optimization considers a rich solution space. (i) In the logical layer, we consider a variety of Global, Local, and Unified network topologies. (ii) In the physical layer, we incorporate well-calibrated microbump overhead models into our area and cost models. We further consider the finite density of microbumps per unit die area,

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	and assess achievable physical wiring distances (hence, achievable link latencies). (iii) In the circuit layer, we explore repeaterless non-pipelined, repeaterless pipelined, repeatered pipelined, and repeatered non-pipelined types of inter-chiplet links. We further consider a gas station link design to enable pipelining in passive inter-chiplet links. • Our heuristic-based cross-layer co-optimization has several novel elements. (i) For a given chiplet placement and network topology, we perform routing and microbump assignment using a flow-based mixed integer-linear program (MILP) to minimize the maximum link latency. (ii) We use workload and network throughput-aware thermal simulation outputs from HotSpot [32] to assess the thermal feasibility of placement and network topology solutions. (iii) We apply simulated annealing to search over our high-dimensional system solution space.”). Coskun at *2 (§ 2) (“In the logical layer, Kannan et al. [16] have evaluated various logical topologies for 2.5D systems, but their work does not consider microbump area overheads, different inter-chiplet link options, or physical implementation of the 2.5D layout. Ahmed et al. [2] propose a hierarchical mesh network for inter-chiplet communication. Both Kannan et al. [16] and Ahmed et al. [2] assume a “minimally active” interposer, which could be unrealistic from a cost perspective (see Section 3). Akgun et al. [3] evaluate three specialized memory-to-core network topologies, yet the evaluation is limited to the logical layer with a static placement of chiplets, and implications of design choices on physical and circuit layers are not explored. None of these works take thermal effects into account or perform a physical design optimization of the 2.5D inter-chiplet network.”). Coskun at *2 (§ 2) (“In the physical layer, Funke et al. [14] have proposed various algorithms that exhaustively search for optimal placement and routing solutions for up to six chiplets. The recent work of Osmonolovskyi et al. [24] handles up to 11-chiplet design complexity using pruning methods. Minz et al. [22] and Fang et al. [13] focus on routing of inter-chiplet links on an interposer. Liu et al. [20] aim to reduce the number of metal layers in the interposer. These works do not consider thermal effects while finding placement and routing solutions. In our

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	prior work [12], we propose a thermally-aware chiplet placement solution. However, our prior work does not perform routing and only computes a placement solution.”). Coskun at *2 (§ 2) (“In the circuit layer, research has generally focused on per-link analyses and optimizations, without considering overheads or tradeoffs with respect to network or system throughput. The works of Stow et al. [28] and Karim et al. [17] explore both repeaterless and repeatered electrical links, while Shamim et al. [27] and Grani et al. [15] respectively consider wireless and photonic links. Ehrett et al. [11] analyze the power and delay overhead of microbumps and conclude that microbump overheads are small. However, they overlook electrostatic discharge (ESD) capacitance, which leads to underestimation network of power and latency.”). Coskun at *2 (§ 2) (“In contrast to these previous works, our methodology jointly considers logical, physical, and circuit design of the inter-chiplet network. We evaluate a variety of logical topologies, while being aware of the network design feasibility in both the circuit layer and the physical layer. In the circuit layer, we evaluate various link design options. In the physical layer, we develop a thermally-aware placement and routing solution. Our cross-layer methodology, thus, obtains 2.5D system solutions that, having more complete and accurate modeling foundations, come closer to defining the true envelope of 2.5D system performance and cost under power and thermal constraints.”). Coskun at *2 (§ 3) (“A cross-layer inter-chiplet network design methodology must comprehend a vast design space that spans the logical, physical and circuit design layers. In this section, we describe the design space for each of these three layers, along with key parameters of interest.”). Coskun at *2-*3 (§ 3.1) (“Each core, together with its L1 and L2 caches, has a square layout. Following our priorwork [12],we study chiplet-based integration of 16 identical chiplets on an interposer, where each chiplet contains 16 cores. Figure 1 illustrates the cross-section view of a 2.5D system. We assume that the 22nm chiplets are placed on an interposer that is designed

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	in 65nm process technology. Microbumps connect the chiplets to the interposer substrate. The system is placed on a System-in-Package (SiP) substrate, with C4 (“flip chip bumps”) connecting the interposer to the SiP substrate. We enable direct comparison of this work with our prior work [12] by designing a Unified-Mesh network using our cross-layer methodology. It should be noted that our broader conclusions are agnostic of the specific core count, core architecture, and technology nodes for the chiplets and the interposer.”). Coskun at *3 (§ 3.2) (“In the logical layer, we explore several different network topologies [35]. We limit the intra-chiplet network to Local-Mesh and Local-Cmesh topologies. For the inter-chiplet network, we design and evaluate Global-Butterfly, Global-Butterdonut [16], and Global-Mesh topologies. For the Unified networks, we evaluate Unified-Mesh and Unified-Cmesh.”). Coskun at *3 (§ 3.3) (“Physical design of the inter-chiplet network consists of placement of the chiplets, along with a routing solution connecting the chiplets that is consistent with the chosen network topology (see Section 3.2). The placement of chiplets affects the temperature map and the length of the links among chiplets, while the routing solution in turn affects the microbump assignment and circuit choices for the link. Further, we explicitly account for the area overhead of microbumps and the associated inter-chiplet drivers and receivers placed along peripheral regions of the chiplets.” (footnote omitted)). Coskun at *3 (§ 3.3) (“Inter-chiplet links can be routed on a passive or an active interposer. Microbumps and ESD protection are required at the beginning and the end of links that go through interposers, and this design constraint adds capacitance [17]. Passive interposers cost less due to their lower manufacturing cost and higher yield [25]. Active interposers allow for repeaters and/or flip-flops (for pipelining) on the interposer. This enables better link bandwidth and latency at the expense of higher manufacturing cost [25]. We conduct a preliminary study of the performance benefit of an active interposer. We observe 2× to 3× latency improvements for the same link length and 50% longer links for the same throughput, but this comes at a 10× cost overhead (\$500 per wafer for passive interposer vs. \$5000 per wafer for active interposer

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	[25]). Given this cost overhead, we rule out active interposers as a realistic option in the near term, and do not consider this option in our present study.”). Coskun at *3 (§ 3.3) (“Passive interposers limit the bandwidth of the signal by degrading rise/fall times. Hence, we use a gas station link, where we can “refuel” a passive link using repeaters and/or flip-flops that are inside other chiplets along the way from the source chiplet to the sink chiplet. Figure 2 shows two implementation schemes for a chiplet-to-chiplet link. Figure 2(a) shows the top view of the paths for the two links connecting Chiplet #1 to Chiplet #3, which are far (e.g., > 10mm) from each other. Figure 2(b) shows a cross-sectional view of the two paths between Chiplet #1 and Chiplet #3. Path 1 uses Chiplet #2 as a gas station, while Path 2 is a direct connection without any gas station. It is important to note the differences between an inter-chiplet repeaterless pipelined link and a gas station link. (i) Pipelining repeaterless links requires an active interposer, while for gas station links we can use a passive interposer. (ii) Active elements required for repeaterless pipelined links are designed using the active interposer’s technology node, while active elements required for gas station links are designed using the chiplet’s technology node. (iii) Using gas station links requires additional microbumps, and in turn, has an area overhead.”). Coskun at *3 (§ 3.3) (“When considering 2.5D inter-chiplet links, recent works have overlooked the microbump overhead while assessing 2.5D integration benefits. Generally, the number of required microbumps will change according to the network topology. An increase in the number of inter-chiplet links increases the number of required microbumps. Further, additional microbumps (20% according to Radojcic [26]) must be reserved for power delivery and signal shielding purposes. Figure 3 shows the chiplet without and with the extra area required for microbumps. Table 1 presents the overhead due to microbumps for different network topologies designed using repeaterless non-pipelined links. The calculations are for the 256-core system divided across 16 chiplets, with each chiplet having an area of 4.5mm × 4.5mm, and a microbump pitch of 45um. Here, h indicates the width of the extra space along the chiplet periphery required for the microbumps used for the inter-chiplet links [26]. The use of gas station link design will further increase microbump count. We do not list the

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	microbump area overhead associated with use of gas station links since this depends upon the placement solution as well as the network type.”). Coskun at *3 (§ 3.4) (“There are multiple circuit design options for inter-chiplet links. For passive interposers, the link on the interposer itself is repeaterless, but with the inclusion of gas stations, the link can use repeaters and/or flip-flops (for pipelining) in intermediate chiplets to regenerate and retime the signal. We limit t_{rise}/t_{cycle} to less than 0.5, to ensure full voltage swing at all nodes in the presence of non-idealities such as supply noise and jitter. We also explore t_{rise}/t_{cycle} of 0.8 that allows us to go longer distances without repeaters. Relaxing the clock period or allowing for multi-cycle bit-periods permits us to use longer inter-chiplet links.”). Coskun at *4 (§ 3.4) (“Figure 4 shows distributed circuit models for link types; (a) repeaterless link in passive interposer, and (b) gas station link in passive interposer. We model wire parasitics using a distributed, multi-segment π model. We use 22nm technology parameters for intra-chiplet components (drivers, receivers, repeaters, and flipflops of the links), while we use 65nm parameters for the interchiplet components of the links. Table 2 shows technology parameter values used in our experiments. We calculate capacitance and resistance based on the model in Wong et al. [30], and we calibrate our stage and path delay estimates based on extraction from layout and Synopsys PrimeTime timing reports.”). Coskun at *4 (§ 4) (“In this section, we describe how we optimize the network design across the layers described in Section 3, using a cross-layer approach. We show our evaluation framework in Figure 5. We first construct oracles for system performance, cost, and interconnect performance. Each of these oracles gives us an element (performance, cost, and latency) of the co-optimization function. Our method for finding a placement solution of chiplets uses a simulated annealing algorithm. We build a search and sort engine that places the oracles and the placement algorithm in a loop to search for a solution across the logical, physical and circuit layers. Table 4 shows the notations we use in the various steps of our crosslayer co-optimization methodology. The placement algorithm uses HotSpot to determine

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	the thermal profile and an MILP is used to find the optimal routing solution. Thus, we determine the feasibility of each placement using HotSpot simulations and the MILP solution.”). Coskun at *4 (§ 4.1) (“We build a system performance oracle that tells us the overall system performance and total core power for a given network topology, voltage-frequency setting, and link latency. To create the oracle, we use Sniper [7] to precompute system performance for a variety of network topologies, voltage-frequency settings, and link latencies. Our system architecture is the 256-core architecture described in Section 3.1. Eight memory controllers are placed next to the top and bottom rows of cores. We implement the inter-chiplet and intra-chiplet network models discussed in Section 3.2 using either passive links or gas station links (see Section 3.3). For passive links without gas stations, we vary inter-chiplet latency values from 1 to 5 cycles, and for gas station links we consider 2- or 3-stage pipelined links. We apply three voltage-frequency settings, (0.9V, 1000MHz), (0.89V, 800MHz) and (0.71V, 533MHz). We fast-forward sequential initialization regions and simulate up to 10 billion instructions in the region of interest using Sniper, with all 256 cores active, to collect performance statistics for five benchmarks. This takes 1.7k CPU hours. We use McPAT [19] to convert the performance results to power traces needed for generating the thermal profile.”). Coskun at *4 (§ 4.2) (“We build a cost oracle that tells us the manufacturing cost of 2.5D systems for a given interposer size, network topology and gas station stage count. We adopt the 2.5D cost model proposed by Stow et al. [28], which takes the cost and yield of chiplets, interposer, and microbump bonding into account, assuming known good dies. We compute the cost of various interposer sizes from 20mm × 20mm to 50mm × 50mm. To estimate the chiplet cost, we compute the number of microbumps required for different network topologies and gas station stages, determine the corresponding chiplet area overhead, and from all these we then calculate the manufacturing cost.” (footnote omitted)). Coskun at *4 (§ 4.3) (“We construct an interconnect performance oracle that tells us the maximum length a signal can travel for a given voltage-frequency setting, rise time constraint,

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	and number of cycles. The link models discussed in Section 3.4 are simulated in HSPICE [21]. For the wire dimensions in the 65nm interposer, i.e., 1μm wire width, 2μm wire pitch, and 1.5μm wire height, the wire resistance is $14.666 \times 10^{-3} \Omega/\mu\text{m}$ and the wire capacitance is $114.726 \times 10^{-3} \text{ fF}/\mu\text{m}$. We use a maximum driver size of 100× the minimum size because the wire latency is largely wire dominated and increasing the driver size beyond 100× does not give latency improvements. We then use these values with our MILP placement solutions to check for placement feasibility. In Table 3, we provide the maximum link lengths we are capable of driving for different cycle numbers, voltage-frequency settings, and rise time constraints. We use power values from HSPICE, along with utilization values from the system performance oracle, to find the total power of the network.”). Coskun at *5 (§ 4.4) (“We use simulated annealing to find a placement that meets the thermal constraint and the maximum link length constraint as evaluated by HotSpot and the routing MILP (with the maximum values provided by the interconnect performance oracle), respectively. We assume a symmetric layout similar to that used in our prior work [12]. As shown in Figure 3(c), we use {s1, s2, s3} as the spacings between chiplets. Simulated annealing searches the solution space in the manner shown in Figure 5. The placement optimization also estimates the microbump area overhead based on the routing solution, link type, and network choice.”). Coskun at *5 (§ 4.4.1) (“We model the 2.5D system in HotSpot using the heterogeneous detailed 3D modeling features [34]. In our thermal model, we use the 2.5D system properties (layer thickness, materials, dimensions of bumps and TSVs, etc.) given in recent work [9, 23]. We use a method similar to our prior work [12], and model each layer of material with separate floorplans on a 64 × 64 grid with ambient temperature at 45°C with default HotSpot sizing convention of the heat sink and spreader. We model leakage as a linear model and assume it to be 30% of the total power at 60°C [33] and rerun HotSpot until temperature convergence is achieved.”).

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	Coskun at *5 (§ 4.4.2) (“We build an MILP that takes the placement of chiplets and the logical network connections as input, and provides the optimal routing solution, including microbump assignment, as an output. The routing optimization is performed internally in the placement optimization as seen in Figure 5. The objective of the MILP is a weighted function of the maximum length of a route on the interposer and the total routing area overhead. We frame the delivery of required numbers of wires between chiplets as multi-commodity flow, and formulate an MILP to find optimal routing solutions that comprehend the finite availability of microbumps in regions of the chiplet periphery.”). Coskun at *5-*6 (§ 4.4.2) (“Based on the inputs to the routing optimization step (see Table 5), we precompute d_{ihjk}, the routing distance (assuming Manhattan routing) from pin clump h on chiplet i to pin clump k on chiplet j, using Equation (1). Equation (2) is the objective function for the MILP that includes the maximum length L, and the total length of the routes. In all reported experiments, we set $\alpha = 1$ and $\beta = 0$. Equation (3) imposes an upper bound on L, ensuring that the solution has routes satisfying the input maximum-length constraint $D_{\max}$. Equation (4) ensures that the flow variable $f^{n_{ihjk}}$ is a nonnegative number. Equation (5) is the flow constraint governing the flow variables $f^{n_{ihjk}}$. It ensures the sum of all flows for a net n, over all pin clumps from chiplet s_n to chiplet t_n, meets the R_{ij} requirement. It also ensures that net flow is 0 for all other (nonsource, non-sink) chiplets for the given net. Equation (6) ensures that there is no input flow (for net n) for the source chiplet of net n. Similarly, Equation (7) ensures that there is no output flow from the sink chiplet of net n. Equation (8) ensures that the sum of input and output flows from a given pin clump is always less than or equal to the capacity of the pin clump. This ensures that all routes have available pins. Equation (9) defines $\lambda^{n_{ihjk}}$ as a boolean value based on $f^{n_{ihjk}}$. This helps identify the maximum route length L, as shown in Equation (10). Equation (11) constrains the maximum number of segments ($S_{\max}$) to be either 1, 2 or 3. If $S_{\max} = 1$, no gas stations are permitted, while if $S_{\max} = 2$ or $S_{\max} = 3$, then gas stations are permitted, allowing for 1 or 2 gas station hops, respectively.”). Coskun at *7 (§ 5) (“In this section, we discuss the results of application of our proposed cross-layer co-optimization methodology. We run multithreaded workloads from SPLASH-2

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	(cholesky, lu.cont) [31], PARSEC (blackscholes, streamcluster) [4], and UHPC (shock) [5] to get a variety of power and performance profiles. For each benchmark, we determine the chiplet placement solution, network routing solution, link type, voltage-frequency setting and network topology. In Figure 6, we show the maximum achievable performance and the corresponding cost of all networks across the five benchmarks for $t_{rise}/t_{cycle} = 0.5$. We show results with and without gas stations.”). Coskun at *7 (§ 5) (“If we do not use gas station links, Unified-Mesh outperforms other networks when running cholesky and streamcluster by 1% to 39%. Unified-Cmesh outperforms all other networks for the remaining benchmarks by <1% to 85%. The higher performance of Unified-Mesh/Cmesh is because they have shorter inter-chiplet links and so they easily achieve single-cycle latency even without gas stations. The latency penalty of long links in Global-Butterfly-Local-Mesh/Cmesh and Global-Butterdonut-Local-Mesh/Cmesh leads to lower performance. On average Unified-Cmesh network has the best performance among all networks. It has more inter-chiplet channels compared to Global networks that results in less contention in the inter-chiplet links, and at the same time it has lower hop count than Unified-Mesh that results in lower latency. The higher performance of Unified-Mesh/Cmesh comes at a cost. Unified-Mesh network is the most expensive and has a manufacturing cost that is 6% to 90% higher than other networks.”). Coskun at *7 (§ 5) (“With gas stations, we can pipeline longer links to improve network throughput. As a result, Global-Butterfly-Local-Mesh/Cmesh and Global-Butterdonut-Local-Mesh/Cmesh networks can achieve better performance with gas stations. Across all benchmarks we see Unified-Cmesh outperforms all other networks by <1% to 21%. However, Unified-Mesh has 1% to 60% higher manufacturing cost compared to all other networks for all benchmarks, except shock. For shock, Global-Butterdonut-Local-Mesh/Cmesh has the highest cost, which is 1% to 20% higher than all remaining networks.”). Coskun at *7 (§ 5) (“To better understand the design space, we also evaluate maximum performance and corresponding cost for networks with and without gas station links when t_{rise}

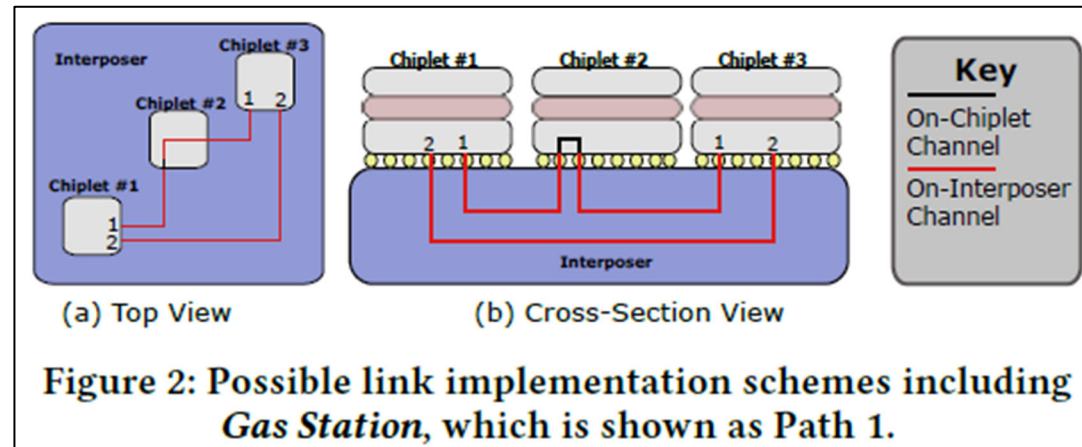
U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	t_{cycle} is 0.8. With this $t_{\text{rise}}/t_{\text{cycle}}$, longer inter-chiplet link lengths without gas stations are feasible. The relaxed length constraint also reduces the microbump and pipeline stage count, which reduces the cost. For $t_{\text{rise}}/t_{\text{cycle}}$ of 0.8, without gas stations, Unified-Cmesh outperforms other networks by <1% to 47%. Unified-Mesh has the highest cost and it is 4% to 52% greater than that of other networks. With gas stations, the performance of Unified-Cmesh is <1% to 11% greater than other networks. Unified-Mesh has the highest cost for all benchmarks except blackscholes and shock, and it is 8% to 60% higher than the cost of the remaining networks. For blackscholes, Global-Butterfly-Local-Cmesh has the highest cost. Here the cost is 2% to 18% higher than the remaining networks. For shock, Global-Butterdonut-Local-Cmesh has the highest cost and it is 2% to 20% higher than the remaining networks.”). Coskun at *7 (§ 5) (“We now highlight differences between outcomes of our previous approach [12] and our present approach. Figure 7(a) shows the placement solution for the cholesky benchmark using our previous approach [12]. That work had predicted a performance boost of 80% with cost comparable to a 2D baseline, while optimizing performance. To make a fair comparison, we apply our cross-layer co-optimization algorithm, running the same benchmark and using the same Unified-Mesh network. Figure 7(b) shows the placement solution from our cross-layer co-optimization. Cost is almost $1.7\times$ higher than that predicted previously [12], while achieving the same (80%) improvement over the 2D baseline system. Figure 7(c) shows the system organization when using our cross-layer cooptimization such that the cost does not exceed the cost of the optimal system organization in Figure 7(a) [12]. Here, we obtain substantially muted performance benefits: rather than 80% performance boost, we achieve a performance boost of 25% compared to the 2D baseline system. Figure 7(d) shows the solution when considering different network topology options while using the cross-layer co-optimization approach to minimize manufacturing cost at equal or higher performance than that of the solution in Figure 7(a). The cost of the solution shown in Figure 7(d) is $1.4\times$ higher than that of the solution in Figure 7(a), but it is 20% lower compared to the solution in Figure 7(b). This 20% cost improvement is achieved due to the choice of Global-Mesh-Local-Cmesh in place of Unified-Mesh. Finally, in Figure 7(e), we show the solution using our cross-layer co-optimization methodology when using all possible design knobs. With Unified-

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	Cmesh, (0.9V, 1000MHz) voltage-frequency setting, and 48mm interposer width, (i) we obtain 90% performance improvement compared to the 2D system, which is 60% better than the performance improvement determined by our prior work; and (ii) we obtain this performance improvement at 16% lower cost compared to our prior work.”). Coskun at *8 (§ 5) (“Figure 8 provides insights regarding the maximum performance possible in a low-cost regime. We sort the 8800 table entries mentioned earlier by manufacturing cost from low to high. We then pick the first 35% of the table entries and identify the placement and routing solution for each network that gives highest performance. With low cost budgets, we see that the higher-performance configurations are dominated by Global-Mesh-Local-Mesh/Cmesh networks. Global-Mesh-Local-Cmesh performs the best in cholesky, lu.cont, and shock, with 1% to 42% better performance than other networks. Global-Mesh-Local-Mesh performs 7% to 50% better than other networks for blackscholes, while Global-Butterfly-Local-Cmesh gives between 1% to 29% better performance than other networks for streamcluster. This is expected, as mesh-like networks have shorter links and can achieve relatively high performance without having to utilize expensive gas station links. Further, in the low-cost regime, we see that Unified-Mesh is not feasible to implement due to the large number of links, which need a large number of microbumps and consequently have a high cost. Since our prior work [12] only considers Unified-Mesh topology, this result shows that it is not a viable solution for low-cost budgets. When we include solutions with up to the 65th cost percentile, we see that Global-Butterdonut-Local-Mesh/Cmesh and Global-Butterfly-Local-Mesh/Cmesh topologies begin to catch up in performance with Global-Mesh-Local-Mesh/Cmesh networks. This is because we can utilize gas station links for the Global-Butterdonut-Local-Mesh/Cmesh and Global-Butterfly-Local-Mesh/Cmesh networks. Global-Mesh-Local-Mesh/Cmesh networks do not benefit as much from the relaxed cost constraint.”). Coskun at *8 (§ 5) (“Finally, we discuss the power of the inter-chiplet network. We see that the highest inter-chiplet network utilization is seen when we run shock on a Unified-Cmesh network. While running shock on Unified-Cmesh, inter-chiplet network power is at most 2% of the overall system power. Theoretically, in very highly threaded applications of the future,

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	we could get much higher network utilizations and then the power of the inter-chiplet network would become a concern.” (footnote omitted)). Coskun at *8 (§ 6) (“In this paper, we have introduced a cross-layer co-optimization methodology for inter-chiplet network design and chiplet placement in 2.5D systems. We have jointly considered network design in the logical, physical, and circuit layers to determine the optimal network choices, link choices, chiplet placements, and link routes to achieve a multi-objective co-optimization goal. We have also proposed to use a gas station link design to enable pipelined interchiplet links when using a passive cost-effective interposer. Our optimization has leveraged well-calibrated models of prior work. We have demonstrated that, compared to 2D systems, our optimized 2.5D systems can achieve 29% better performance with the same manufacturing cost, or 25% lower cost with the same performance.”). Coskun at *8 (§ 6) (“Throughout this work, we have focused on running a single parallel application at a time and have shown the co-optimization outcomes for a variety of benchmarks. Based on these results, a 2.5D system can be further optimized in an application-aware manner (e.g., based on specific applications or worst-/average-case results). Interesting open problems include co-optimization with multi-application scenarios, allocation of threads in a network-aware manner, co-optimization with heterogeneous chiplets, and exploration of active interposer. Also, while we have designed our system for the worst-case link latencies under a global latency constraint, future work involves designing networks with variable link latencies.”). This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA’s other invalidity charts and cover pleading.

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
1[D]: wherein the passive crosslink is dedicated for inter-chiplet communications.	Including at least under Complainant's apparent infringement theory, Coskun discloses and/or renders obvious this limitation. <i>See</i> evidence cited for '803 patent limitations 1[A]-1[C]. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
2. The system of claim 1, wherein the passive crosslink comprises a passive interposer die.	Including at least under Complainant's apparent infringement theory, Coskun discloses and/or renders obvious this claim. <i>See, e.g.,</i> Coskun at Fig. 1. Figure 1: The cross-section of a 2.5D system.

Coskun at Fig. 2.



Coskun at *1 (§ 1) (“The need to sustain the historical performance and cost scaling in computing systems has led to a growing interest in 2.5D systems [1, 8, 9, 15, 16, 29]. In 2.5D design, multiple chiplets are placed on a silicon interposer, and the chiplets communicate using links integrated into the interposer. 2.5D integration technology provides multiple potential benefits compared to 2D systems, including greater system performance within thermal constraints [12], heterogeneous integration of multiple technologies [1, 6], and reduced overall system cost [16]. However, 2.5D integration technology also opens up a number of design challenges, ranging from circuit and physical challenges (design and routing of inter-chiplet links, placement and floorplanning of chiplets on the interposer, microbump assignment, etc.) to architectural and system-level challenges (design of the inter-chiplet network architecture, partitioning a system into 2.5D-integrated heterogeneous functional components, etc.). Many other technical and business challenges, including design for thermomechanical stress, test strategy, and supply chain structure, are identified by Radojcic [26].”).

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	Coskun at *1 (§ 1) (“In this paper, we perform a cross-layer co-optimization of 2.5D inter-chiplet network design and chiplet placement. Our cooptimization methodology focuses on network topologies, link circuit options, and microbump pitch- and interconnect RC-aware routing of links. It maximizes performance and/or minimizes cost at the system level, while satisfying system power and thermal constraints. The need for such a cross-layer methodology as ours can be easily seen by considering the following. If we adopt a topdown approach, an architecture analysis of network topologies tells us that high-radix, low-diameter networks should be used for inter-chiplet networks, as they provide the best overall system performance (in instructions per cycle). However, in the physical layer, realization of high-radix, low-diameter networks requires long wires, which can limit the network performance and, hence, the overall system performance. Using repeaters on long wires to improve performance would necessitate active (rather than passive) interposer technology. Since active interposers are 10× more expensive than passive interposers [25], the system cost equation changes and the top-down intuition is flawed. On the other hand, a bottom-up, cost-centric perspective prompts the use of a passive interposer, which can only support repeaterless links in the circuit layer, thus limiting link performance and maximum link length. This leads to the adoption of low-radix, high-diameter networks at the inter-chiplet level, which lowers overall system performance.”). Coskun at *2 (§ 1) (“The main contributions of this paper are as follows: • We develop a cross-layer co-optimization methodology that optimizes inter-chiplet network design jointly with chiplet placement across logical, physical, and circuit layers. Our methodology optimizes a given 2.5D system for performance, cost, and wirelength, while ensuring that it is thermally safe. The outcome of the co-optimization comprises placement of chiplets on the interposer, logical topology of the interchiplet network, and circuit design and routing of the links that form the network. • Our co-optimization considers a rich solution space. (i) In the logical layer, we consider a variety of Global, Local, and Unified network topologies. (ii) In the physical layer, we incorporate well-calibrated microbump overhead models into our area and cost models. We further consider the finite density of microbumps per unit die area,

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	and assess achievable physical wiring distances (hence, achievable link latencies). (iii) In the circuit layer, we explore repeaterless non-pipelined, repeaterless pipelined, repeatered pipelined, and repeatered non-pipelined types of inter-chiplet links. We further consider a gas station link design to enable pipelining in passive inter-chiplet links. • Our heuristic-based cross-layer co-optimization has several novel elements. (i) For a given chiplet placement and network topology, we perform routing and microbump assignment using a flow-based mixed integer-linear program (MILP) to minimize the maximum link latency. (ii) We use workload and network throughput-aware thermal simulation outputs from HotSpot [32] to assess the thermal feasibility of placement and network topology solutions. (iii) We apply simulated annealing to search over our high-dimensional system solution space.”). Coskun at *2 (§ 2) (“In the logical layer, Kannan et al. [16] have evaluated various logical topologies for 2.5D systems, but their work does not consider microbump area overheads, different inter-chiplet link options, or physical implementation of the 2.5D layout. Ahmed et al. [2] propose a hierarchical mesh network for inter-chiplet communication. Both Kannan et al. [16] and Ahmed et al. [2] assume a “minimally active” interposer, which could be unrealistic from a cost perspective (see Section 3). Akgun et al. [3] evaluate three specialized memory-to-core network topologies, yet the evaluation is limited to the logical layer with a static placement of chiplets, and implications of design choices on physical and circuit layers are not explored. None of these works take thermal effects into account or perform a physical design optimization of the 2.5D inter-chiplet network.”). Coskun at *2 (§ 2) (“In the physical layer, Funke et al. [14] have proposed various algorithms that exhaustively search for optimal placement and routing solutions for up to six chiplets. The recent work of Osmonolovskyi et al. [24] handles up to 11-chiplet design complexity using pruning methods. Minz et al. [22] and Fang et al. [13] focus on routing of inter-chiplet links on an interposer. Liu et al. [20] aim to reduce the number of metal layers in the interposer. These works do not consider thermal effects while finding placement and routing solutions. In our

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	prior work [12], we propose a thermally-aware chiplet placement solution. However, our prior work does not perform routing and only computes a placement solution.”). Coskun at *3 (§ 3.3) (“Inter-chiplet links can be routed on a passive or an active interposer. Microbumps and ESD protection are required at the beginning and the end of links that go through interposers, and this design constraint adds capacitance [17]. Passive interposers cost less due to their lower manufacturing cost and higher yield [25]. Active interposers allow for repeaters and/or flip-flops (for pipelining) on the interposer. This enables better link bandwidth and latency at the expense of higher manufacturing cost [25]. We conduct a preliminary study of the performance benefit of an active interposer. We observe 2× to 3× latency improvements for the same link length and 50% longer links for the same throughput, but this comes at a 10× cost overhead (\$500 per wafer for passive interposer vs. \$5000 per wafer for active interposer [25]). Given this cost overhead, we rule out active interposers as a realistic option in the near term, and do not consider this option in our present study.”). Coskun at *3 (§ 3.3) (“Passive interposers limit the bandwidth of the signal by degrading rise/fall times. Hence, we use a gas station link, where we can “refuel” a passive link using repeaters and/or flip-flops that are inside other chiplets along the way from the source chiplet to the sink chiplet. Figure 2 shows two implementation schemes for a chiplet-to-chiplet link. Figure 2(a) shows the top view of the paths for the two links connecting Chiplet #1 to Chiplet #3, which are far (e.g., > 10mm) from each other. Figure 2(b) shows a cross-sectional view of the two paths between Chiplet #1 and Chiplet #3. Path 1 uses Chiplet #2 as a gas station, while Path 2 is a direct connection without any gas station. It is important to note the differences between an inter-chiplet repeaterless pipelined link and a gas station link. (i) Pipelining repeaterless links requires an active interposer, while for gas station links we can use a passive interposer. (ii) Active elements required for repeaterless pipelined links are designed using the active interposer’s technology node, while active elements required for gas station links are designed using the chiplet’s technology node. (iii) Using gas station links requires additional microbumps, and in turn, has an area overhead.”).

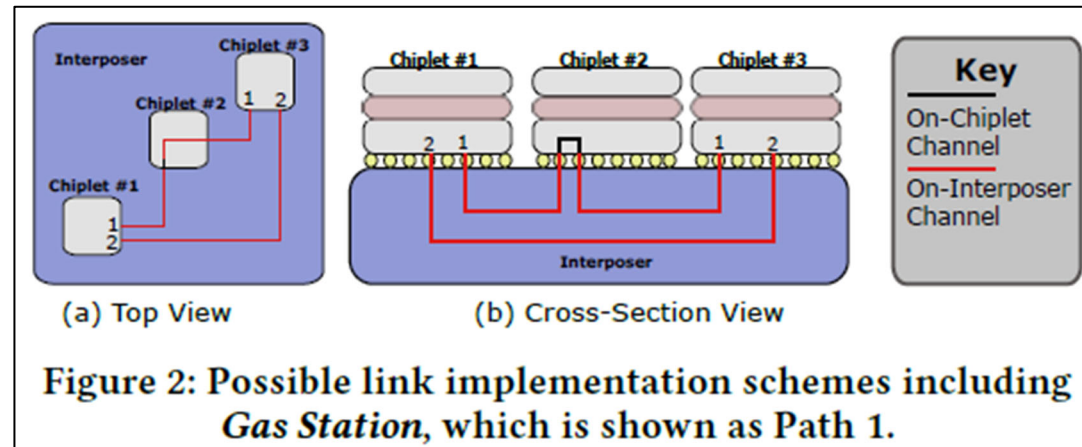
U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	Coskun at *3 (§ 3.4) (“There are multiple circuit design options for inter-chiplet links. For passive interposers, the link on the interposer itself is repeaterless, but with the inclusion of gas stations, the link can use repeaters and/or flip-flops (for pipelining) in intermediate chiplets to regenerate and retime the signal. We limit t_{rise}/t_{cycle} to less than 0.5, to ensure full voltage swing at all nodes in the presence of non-idealities such as supply noise and jitter. We also explore t_{rise}/t_{cycle} of 0.8 that allows us to go longer distances without repeaters. Relaxing the clock period or allowing for multi-cycle bit-periods permits us to use longer inter-chiplet links.”). Coskun at *4 (§ 3.4) (“Figure 4 shows distributed circuit models for link types; (a) repeaterless link in passive interposer, and (b) gas station link in passive interposer. We model wire parasitics using a distributed, multi-segment π model. We use 22nm technology parameters for intra-chiplet components (drivers, receivers, repeaters, and flipflops of the links), while we use 65nm parameters for the interchiplet components of the links. Table 2 shows technology parameter values used in our experiments. We calculate capacitance and resistance based on the model in Wong et al. [30], and we calibrate our stage and path delay estimates based on extraction from layout and Synopsys PrimeTime timing reports.”). Coskun at *4 (§ 4.2) (“We build a cost oracle that tells us the manufacturing cost of 2.5D systems for a given interposer size, network topology and gas station stage count. We adopt the 2.5D cost model proposed by Stow et al. [28], which takes the cost and yield of chiplets, interposer, and microbump bonding into account, assuming known good dies. We compute the cost of various interposer sizes from $20\text{mm} \times 20\text{mm}$ to $50\text{mm} \times 50\text{mm}$. To estimate the chiplet cost, we compute the number of microbumps required for different network topologies and gas station stages, determine the corresponding chiplet area overhead, and from all these we then calculate the manufacturing cost.” (footnote omitted)). Coskun at *4 (§ 4.3) (“We construct an interconnect performance oracle that tells us the maximum length a signal can travel for a given voltage-frequency setting, rise time constraint, and number of cycles. The link models discussed in Section 3.4 are simulated in HSPICE

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	[21]. For the wire dimensions in the 65nm interposer, i.e., 1μm wire width, 2μm wire pitch, and 1.5μm wire height, the wire resistance is $14.666 \times 10^{-3} \Omega/\mu\text{m}$ and the wire capacitance is $114.726 \times 10^{-3} \text{ fF}/\mu\text{m}$. We use a maximum driver size of 100× the minimum size because the wire latency is largely wire dominated and increasing the driver size beyond 100× does not give latency improvements. We then use these values with our MILP placement solutions to check for placement feasibility. In Table 3, we provide the maximum link lengths we are capable of driving for different cycle numbers, voltage-frequency settings, and rise time constraints. We use power values from HSPICE, along with utilization values from the system performance oracle, to find the total power of the network.”). Coskun at *5 (§ 4.4.2) (“We build an MILP that takes the placement of chiplets and the logical network connections as input, and provides the optimal routing solution, including microbump assignment, as an output. The routing optimization is performed internally in the placement optimization as seen in Figure 5. The objective of the MILP is a weighted function of the maximum length of a route on the interposer and the total routing area overhead. We frame the delivery of required numbers of wires between chiplets as multi-commodity flow, and formulate an MILP to find optimal routing solutions that comprehend the finite availability of microbumps in regions of the chiplet periphery.”). Coskun at *7 (§ 5) (“We now highlight differences between outcomes of our previous approach [12] and our present approach. Figure 7(a) shows the placement solution for the cholesky benchmark using our previous approach [12]. That work had predicted a performance boost of 80% with cost comparable to a 2D baseline, while optimizing performance. To make a fair comparison, we apply our cross-layer co-optimization algorithm, running the same benchmark and using the same Unified-Mesh network. Figure 7(b) shows the placement solution from our cross-layer co-optimization. Cost is almost 1.7× higher than that predicted previously [12], while achieving the same (80%) improvement over the 2D baseline system. Figure 7(c) shows the system organization when using our cross-layer cooptimization such that the cost does not exceed the cost of the optimal system organization in Figure 7(a) [12]. Here, we obtain substantially muted performance benefits: rather than 80% performance boost,

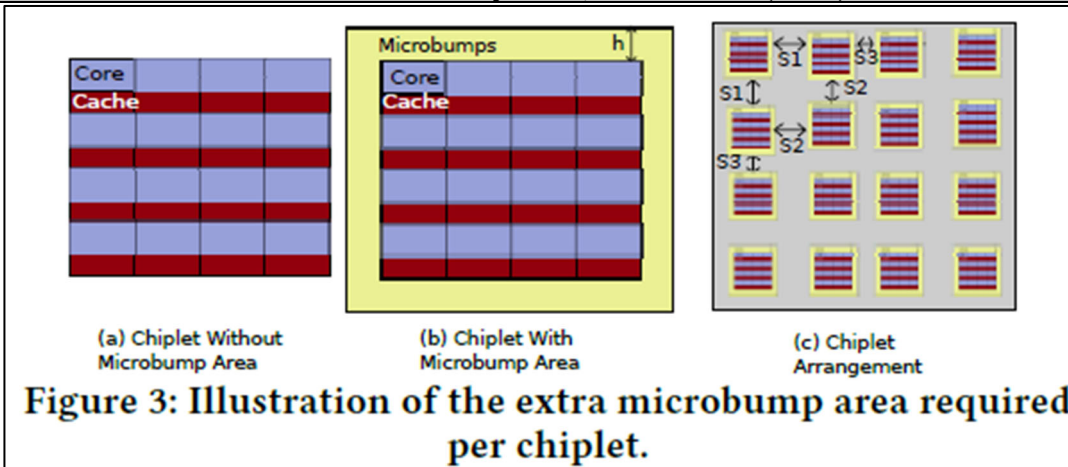
U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	we achieve a performance boost of 25% compared to the 2D baseline system. Figure 7(d) shows the solution when considering different network topology options while using the cross-layer co-optimization approach to minimize manufacturing cost at equal or higher performance than that of the solution in Figure 7(a). The cost of the solution shown in Figure 7(d) is 1.4× higher than that of the solution in Figure 7(a), but it is 20% lower compared to the solution in Figure 7(b). This 20% cost improvement is achieved due to the choice of Global-Mesh-Local-Cmesh in place of Unified-Mesh. Finally, in Figure 7(e), we show the solution using our cross-layer co-optimization methodology when using all possible design knobs. With Unified-Cmesh, (0.9V, 1000MHz) voltage-frequency setting, and 48mm interposer width, (i) we obtain 90% performance improvement compared to the 2D system, which is 60% better than the performance improvement determined by our prior work; and (ii) we obtain this performance improvement at 16% lower cost compared to our prior work.”). Coskun at *8 (§ 6) (“In this paper, we have introduced a cross-layer co-optimization methodology for inter-chiplet network design and chiplet placement in 2.5D systems. We have jointly considered network design in the logical, physical, and circuit layers to determine the optimal network choices, link choices, chiplet placements, and link routes to achieve a multi-objective co-optimization goal. We have also proposed to use a gas station link design to enable pipelined interchiplet links when using a passive cost-effective interposer. Our optimization has leveraged well-calibrated models of prior work. We have demonstrated that, compared to 2D systems, our optimized 2.5D systems can achieve 29% better performance with the same manufacturing cost, or 25% lower cost with the same performance.”). Coskun at *8 (§ 6) (“Throughout this work, we have focused on running a single parallel application at a time and have shown the co-optimization outcomes for a variety of benchmarks. Based on these results, a 2.5D system can be further optimized in an application-aware manner (e.g., based on specific applications or worst-/average-case results). Interesting open problems include co-optimization with multi-application scenarios, allocation of threads in a network-aware manner, co-optimization with heterogeneous chiplets, and exploration of active interposer. Also, while we have designed our system for the worst-case link latencies

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	under a global latency constraint, future work involves designing networks with variable link latencies.”). <i>See also</i> evidence cited for '803 patent limitation 1[C]. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA’s other invalidity charts and cover pleading.
3[A]: The system of claim 1, wherein the first GPU chiplet comprises a first PHY region including conductor structures for chiplet-to-chiplet communications,	Including at least under Complainant’s apparent infringement theory, Coskun discloses and/or renders obvious this limitation. <i>See, e.g.,</i> Coskun at Fig. 1. The diagram illustrates a cross-section of a 2.5D system. At the top, two chiplets, labeled 'Chiplet #1' and 'Chiplet #2', are shown. Below them is a layer of 'Microbumps'. Underneath the microbumps is the 'Frontside Metal Layer'. Below that is the 'Through Silicon Vias' layer, which contains several vertical vias. Below the through silicon vias is the 'Backside Metal Layer'. At the bottom of the stack is the 'SiP Substrate'. 'Flip Chip Bumps' are shown on the bottom surface of the SiP Substrate. The entire structure is labeled 'Figure 1: The cross-section of a 2.5D system.'

Coskun at Fig. 2.



Coskun at Fig. 3.



Coskun et al. (Abstract) (“2.5D integration technology is gaining popularity in the design of homogeneous and heterogeneous many-core computing systems. 2.5D network design, both inter- and intra-chiplet, impacts overall system performance as well as its manufacturing cost and thermal feasibility. This paper introduces a cross-layer methodology for designing networks in 2.5D systems. We optimize the network design and chiplet placement jointly across logical, physical, and circuit layers to achieve an energy-efficient network, while maximizing system performance, minimizing manufacturing cost, and adhering to thermal constraints. In the logical layer, our cooptimization considers eight different network topologies. In the physical layer, we consider routing, microbump assignment, and microbump pitch constraints to account for the extra costs associated with microbump utilization in the inter-chiplet communication. In the circuit layer, we consider both passive and active links with five different link types, including a gas station link design. Using our cross-layer methodology results in more accurate determination of (superior) inter-chiplet network and 2.5D system designs compared to prior methods. Compared to 2D systems, our approach

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	achieves 29% better performance with the same manufacturing cost, or 25% lower cost with the same performance.”). Coskun at *2-*3 (§ 3.1) (“Each core, together with its L1 and L2 caches, has a square layout. Following our priorwork [12],we study chiplet-based integration of 16 identical chiplets on an interposer, where each chiplet contains 16 cores. Figure 1 illustrates the cross-section view of a 2.5D system. We assume that the 22nm chiplets are placed on an interposer that is designed in 65nm process technology. Microbumps connect the chiplets to the interposer substrate. The system is placed on a System-in-Package (SiP) substrate, with C4 (“flip chip bumps”) connecting the interposer to the SiP substrate. We enable direct comparison of this work with our prior work [12] by designing a Unified-Mesh network using our cross-layer methodology. It should be noted that our broader conclusions are agnostic of the specific core count, core architecture, and technology nodes for the chiplets and the interposer.”). Coskun at *3 (§ 3.3) (“Physical design of the inter-chiplet network consists of placement of the chiplets, along with a routing solution connecting the chiplets that is consistent with the chosen network topology (see Section 3.2). The placement of chiplets affects the temperature map and the length of the links among chiplets, while the routing solution in turn affects the microbump assignment and circuit choices for the link. Further, we explicitly account for the area overhead of microbumps and the associated inter-chiplet drivers and receivers placed along peripheral regions of the chiplets.” (footnote omitted)). Coskun at *3 (§ 3.3) (“Inter-chiplet links can be routed on a passive or an active interposer. Microbumps and ESD protection are required at the beginning and the end of links that go through interposers, and this design constraint adds capacitance [17]. Passive interposers cost less due to their lower manufacturing cost and higher yield [25]. Active interposers allow for repeaters and/or flip-flops (for pipelining) on the interposer. This enables better link bandwidth and latency at the expense of higher manufacturing cost [25]. We conduct a preliminary study of the performance benefit of an active interposer. We observe 2× to 3× latency improvements for the same link length and 50% longer links for the same throughput, but this comes at a 10×

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	cost overhead (\$500 per wafer for passive interposer vs. \$5000 per wafer for active interposer [25]). Given this cost overhead, we rule out active interposers as a realistic option in the near term, and do not consider this option in our present study.”). Coskun at *3 (§ 3.3) (“Passive interposers limit the bandwidth of the signal by degrading rise/fall times. Hence, we use a gas station link, where we can “refuel” a passive link using repeaters and/or flip-flops that are inside other chiplets along the way from the source chiplet to the sink chiplet. Figure 2 shows two implementation schemes for a chiplet-to-chiplet link. Figure 2(a) shows the top view of the paths for the two links connecting Chiplet #1 to Chiplet #3, which are far (e.g., > 10mm) from each other. Figure 2(b) shows a cross-sectional view of the two paths between Chiplet #1 and Chiplet #3. Path 1 uses Chiplet #2 as a gas station, while Path 2 is a direct connection without any gas station. It is important to note the differences between an inter-chiplet repeaterless pipelined link and a gas station link. (i) Pipelining repeaterless links requires an active interposer, while for gas station links we can use a passive interposer. (ii) Active elements required for repeaterless pipelined links are designed using the active interposer’s technology node, while active elements required for gas station links are designed using the chiplet’s technology node. (iii) Using gas station links requires additional microbumps, and in turn, has an area overhead.”). Coskun at *3 (§ 3.3) (“When considering 2.5D inter-chiplet links, recent works have overlooked the microbump overhead while assessing 2.5D integration benefits. Generally, the number of required microbumps will change according to the network topology. An increase in the number of inter-chiplet links increases the number of required microbumps. Further, additional microbumps (20% according to Radojcic [26]) must be reserved for power delivery and signal shielding purposes. Figure 3 shows the chiplet without and with the extra area required for microbumps. Table 1 presents the overhead due to microbumps for different network topologies designed using repeaterless non-pipelined links. The calculations are for the 256-core system divided across 16 chiplets, with each chiplet having an area of 4.5mm × 4.5mm, and a microbump pitch of 45um. Here, h indicates the width of the extra space along the chiplet periphery required for the microbumps used for the inter-chiplet links [26]. The use

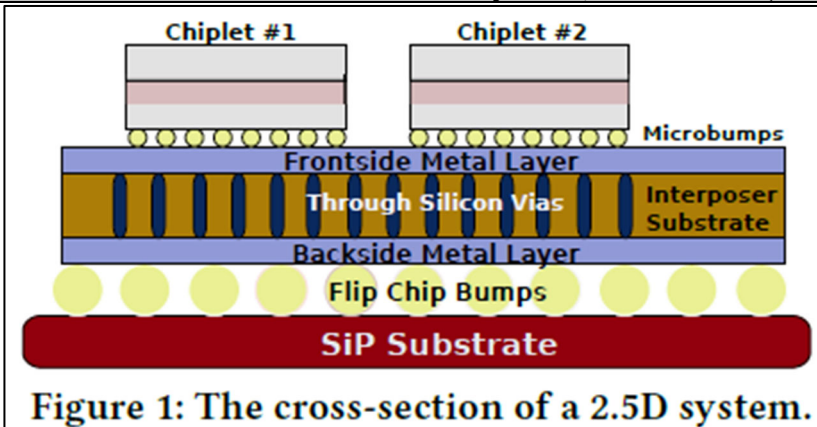
U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	of gas station link design will further increase microbump count. We do not list the microbump area overhead associated with use of gas station links since this depends upon the placement solution as well as the network type.”). <i>See also</i> evidence cited for '803 patent limitation 1[C]. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA’s other invalidity charts and cover pleading.
3[B]: and further wherein the second GPU chiplet comprises a second PHY region including conductor structures for chiplet-to-chiplet communications.	Including at least under Complainant’s apparent infringement theory, Coskun discloses and/or renders obvious this limitation. <i>See</i> evidence cited for '803 patent limitation 3[A]. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA’s other invalidity charts and cover pleading.
7. The system of claim 1, wherein the passive crosslink communicably couples all GPU chiplets in the GPU chiplet array.	Including at least under Complainant’s apparent infringement theory, Coskun discloses and/or renders obvious this claim. <i>See</i> evidence cited for '803 patent limitation 1[D]. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
8[A]: The system of claim 1, further comprising: a first cache memory hierarchy at the first GPU chiplet,	Including at least under Complainant's apparent infringement theory, Coskun discloses and/or renders obvious this limitation. See, e.g., Coskun at Fig. 3. (a) Chiplet Without Microbump Area (b) Chiplet With Microbump Area (c) Chiplet Arrangement Figure 3: Illustration of the extra microbump area required per chiplet. Coskun at *2 (§ 3.1) (“Our studies use a 256-core homogeneous (i.e., all cores are of the same type) system. To enable comparisons against the previous literature, we specifically adopt the core design used in our prior work [12]. Cores have the following architectural specifications: • 16KB I/D L1 Cache • 256KB Private L2 Cache • 0.93mm² Core + L1 Area

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	• 0.35mm² L2 Area • 1.28mm² (1.13mm × 1.13mm) Total Area [33] • 18mm × 18mm Total 256-core Chip Area”). Coskun at *2-*3 (§ 3.1) (“Each core, together with its L1 and L2 caches, has a square layout. Following our priorwork [12],we study chiplet-based integration of 16 identical chiplets on an interposer, where each chiplet contains 16 cores. Figure 1 illustrates the cross-section view of a 2.5D system. We assume that the 22nm chiplets are placed on an interposer that is designed in 65nm process technology. Microbumps connect the chiplets to the interposer substrate. The system is placed on a System-in-Package (SiP) substrate, with C4 (“flip chip bumps”) connecting the interposer to the SiP substrate. We enable direct comparison of this work with our prior work [12] by designing a Unified-Mesh network using our cross-layer methodology. It should be noted that our broader conclusions are agnostic of the specific core count, core architecture, and technology nodes for the chiplets and the interposer.”). <i>See also</i> evidence cited for ’803 patent limitation 1[C]. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA’s other invalidity charts and cover pleading.
8[B]: wherein a first level of the first cache memory hierarchy is coherent within the first GPU chiplet; and	Including at least under Complainant’s apparent infringement theory, Coskun discloses and/or renders obvious this limitation. <i>See</i> evidence cited for ’803 patent limitation 8[A]. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
8[C]: a second cache memory hierarchy at the second GPU chiplet,	Including at least under Complainant's apparent infringement theory, Coskun discloses and/or renders obvious this limitation. <i>See</i> evidence cited for '803 patent limitation 8[A]. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
8[D]: wherein a first level of the second cache memory hierarchy is coherent within the second GPU chiplet.	Including at least under Complainant's apparent infringement theory, Coskun discloses and/or renders obvious this limitation. <i>See</i> evidence cited for '803 patent limitation 8[A]. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
9[A]: The system of claim 8, further comprising: a unified cache memory including both a last level of the first cache memory hierarchy and a last level of the second cache memory hierarchy,	Including at least under Complainant's apparent infringement theory, Coskun discloses and/or renders obvious this limitation. <i>See</i> evidence cited for '803 patent claim 8. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of

U.S. Patent No. 11,841,803	Ayse Coskun et al., <i>A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems</i> , ICCAD '18 (2018)
	ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
9[B]: wherein the unified cache memory is coherent across all GPU chiplets of the GPU chiplet array.	Including at least under Complainant's apparent infringement theory, Coskun discloses and/or renders obvious this limitation. <i>See</i> evidence cited for '803 patent claim 8. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
10. The system of claim 1, further comprising: a plurality of conductive pillars coupling a circuit board substrate to a first non-PHY region of the first GPU chiplet and a second non-PHY region of the second GPU chiplet.	Including at least under Complainant's apparent infringement theory, Coskun discloses and/or renders obvious this claim. <i>See, e.g.,</i> Coskun at Fig. 1.



Coskun at *5 (§ 4.4.1) (“We model the 2.5D system in HotSpot using the heterogeneous detailed 3D modeling features [34]. In our thermal model, we use the 2.5D system properties (layer thickness, materials, dimensions of bumps and TSVs, etc.) given in recent work [9, 23]. We use a method similar to our prior work [12], and model each layer of material with separate floorplans on a 64×64 grid with ambient temperature at 45°C with default HotSpot sizing convention of the heat sink and spreader. We model leakage as a linear model and assume it to be 30% of the total power at 60°C [33] and rerun HotSpot until temperature convergence is achieved.”).

See also evidence cited for '803 patent limitation 1[C].

This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Coskun alone; (2) Coskun in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Coskun in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.