# Leveraging Thermally-Aware Chiplet Organization in 2.5D Systems to Reclaim Dark Silicon

Furkan Eris[1], Ajay Joshi[1], Andrew B. Kahng[2,3], Yenai Ma[1], Saiful Mojumder[1] and Tiansheng Zhang[1]
[1]ECE Department, Boston University, Boston, MA, USA; [2]ECE and [3]CSE Departments, UC San Diego, La Jolla, CA, USA
{fe, joshi, yenai, msam, tszhang}@bu.edu, abk@cs.ucsd.edu

***Abstract*—As on-chip power densities of manycore systems continue to increase, one cannot simultaneously run all the cores due to thermal constraints. This phenomenon, known as the 'dark silicon' problem, leads to inactive regions on the chip and limits the performance of manycore systems. This paper proposes to reclaim dark silicon through a thermally-aware chiplet organization technique in 2.5D manycore systems. The proposed technique adjusts the interposer size and the spacing between adjacent chiplets to reduce the peak temperature of the overall system. In this way, a system can operate with a larger number of active cores at a higher frequency without violating thermal constraints, thereby achieving higher performance. To determine the chiplet organization that jointly maximizes performance and minimizes manufacturing cost, we formulate and solve an optimization problem that considers temperature and interposer size constraints of 2.5D systems. We design a multi-start greedy approach to find (near-)optimal solutions efficiently. Our analysis demonstrates that by using our proposed technique, an optimized 2.5D manycore system improves performance by 41% and 16% on average and by up to 87% and 39% for temperature thresholds of 85$^oC$ and 105$^oC$, respectively, compared to a traditional single-chip system at the same manufacturing cost. When maintaining the same performance as an equivalent single-chip system, our approach is able to reduce the 2.5D system manufacturing cost by 36%.**

## I. Introduction

Over the past decade, CMOS technology scaling has slowed down, and as a result, it is difficult to sustain the historic performance improvements in CMOS-based VLSI systems. To address this challenge, the computing industry has moved towards packing an increasing number of cores on a single die and using thread-level parallelism to continuously improve performance. At the same time, the on-chip power density has risen with shrinking transistor feature size. This increasing power density has led to 'dark silicon' [1] on a chip. As a result in manycore systems not all cores can be operated at the highest frequency or even turned on simultaneously due to thermal constraints. Thus, there is a significant amount of performance that is 'left on the table' in today's manycore systems.

A variety of solutions have been proposed to address the dark silicon problem at both hardware level [2]–[5] and system management level [6], [7] for single-chip systems. These techniques help balance the heat dissipation across the chip, thereby improving system energy efficiency under thermal constraints. However, these techniques are not able to maximize the performance in manycore systems.

In tandem with technology scaling and the move to manycore systems, die-stacking technologies such as 2.5D and 3D integration have emerged to improve system performance [8]–[10]. 3D integration, which stacks dies vertically to form a system, reduces system footprint and increases memory bandwidth [9], but exacerbates the thermal issues [8]. 2.5D integration, which integrates small chiplets on a silicon interposer, is less prone to the thermal challenges observed in 3D stacking [10]. Moreover, it provides additional routing resources through the interposer, and is more cost-effective [9], [10]. Currently, 2.5D integration technology is being extensively investigated by both academia and industry [9], [11]–[14].

In 2.5D integration, the general approach to arrange chiplets is to integrate them as close as possible on an interposer to save cost. There is however an opportunity here to solve the 'dark silicon' problem by organizing the chiplets in a thermally-aware fashion such that we can lower the overall manycore system temperature and in turn improve performance (by having more active cores operating at higher frequency) without significantly increasing the cost. In this paper, we first explore the impact of chiplet placement on the cost and thermal behavior of 2.5D manycore systems. We then propose a thermally-aware chiplet organization strategy to address the dark silicon problem in 2.5D manycore systems. Our main contributions are as follows:

- We perform a detailed design space exploration of chiplet organizations in 2.5D manycore systems to analyze the impact of chiplet count, power density, and interposer size on the system temperature.
- We propose to strategically insert spacing between the chiplets of 2.5D manycore systems to lower the system temperature. This reduction enables higher operating frequency and/or more active cores in the 2.5D manycore system under the same temperature threshold, which in turn improves the overall system performance.
- We design a multi-start greedy approach to efficiently find the (near-)optimal thermally-aware chiplet organization that jointly maximizes the manycore system performance and minimizes the overall system manufacturing cost.

## II. Related Work

Over the past few years, a number of solutions have been proposed to alleviate the dark silicon problem. The proposed solutions include the use of specialized cores [3], DVFS [2], near-threshold computing [4], approximate computing [5], power budgeting [6], and computational sprinting [7]. A specialized core enables efficient execution of a specific application with a smaller number of transistors, but it cannot execute other types of applications efficiently. Applying DVFS degrades system performance, while near-threshold and approximate computing trade off accuracy and reliability for energy efficiency. Power budgeting enables operation at a thermally-safe power rather than at a constant thermal design power (TDP) and achieves a higher total performance. Computational sprinting (where the system runs with a larger number of cores in short bursts) incorporates phase-change materials for higher thermal capacitance, and thus allows violation of the thermal power budget for a short time. Power budgeting and computational sprinting, however, require a 'cooling down' period after the performance boost.

A number of previous approaches have introduced thermally-aware floorplanning methods to reduce hot spots [15], to achieve

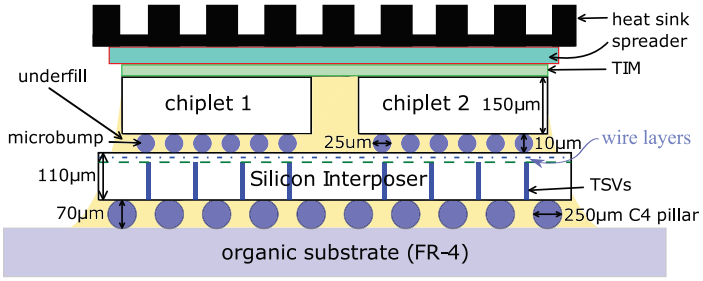


Fig. 1: *Cross-sectional view of a 2.5D system.*

high performance [16], and to reduce peak temperature [17] and thermal gradients [18] of 3D ICs. All of these works consider placement of components to reduce temperature, but they do not focus on 2.5D systems.

In contrast with the prior work, our work leverages thermally-aware 2.5D integration to reclaim dark silicon. We are the first to propose a thermally-aware chiplet organization, where we strategically place chiplets in a thermally-aware fashion to facilitate heat dissipation, and in turn raise the thermally-safe power budget without increasing the cooling cost, and in turn improve performance. Moreover, our approach complements other approaches such as near-threshold computing, approximate computing, and specialized cores to reclaim dark silicon.

## III. Thermally-Aware Chiplet Organization

In this section we present the details of our example 2.5D system, manufacturing cost model and thermal behavior of 2.5D systems, and our approach to optimize the chiplet organization. All notations used in this section are listed in Table II.

### A. 2.5D System Overview

We use a 256-core homogeneous system operating at $1GHz$ as an example manycore system in this work. The core architecture of this system is based on the IA-32 core from the Intel Single-Chip Cloud Computer (SCC) [19], with size and power scaled to $22nm$ technology [20]. Each core has a 16 KB I/D L1 cache and a 256 KB private L2 cache. The area of each core (including L1 cache) is $0.93mm^2$, and the area of each L2 cache is $0.35mm^2$. We assume each L2 cache is placed next to the corresponding core, and each core together with its L2 cache is square shaped, with an area of $1.28mm^2$ ($1.13mm \times 1.13mm$) [20]. The total size of the 256-core single chip is $18mm \times 18mm$. We assume the 256-core system has 8 memory controllers distributed along the two opposite edges of the chip and the DRAM is located off-chip.

In the example 256-core 2.5D system (Fig. 1), we split a single chip into chiplets, place the chiplets onto a $65nm$ passive interposer, and use microbumps connecting the chiplets and the interposer. There are through-silicon vias (TSVs) inside the interposer to connect its upper and lower layers. We place the interposer on top of a substrate using C4 bumps for connection. Epoxy resins are used to underfill the spacing among C4 bumps, the spacing among microbumps, and the empty spaces among chiplets [21]. The dimensions of the 2.5D system (shown in Table I) are based on the prototypes from CEA-Leti [13] and Xilinx [14]. Our evaluation uses the conventional 2D single-chip system as a baseline, where the 256-core chip is placed directly on top of an organic substrate with C4 bumps for connection [21].

We use an electrical mesh network (single-cycle routers and single-cycle links) for the example 256-core system. Intra-chiplet communication is through on-chiplet interconnects, while inter-chiplet communication is through links in the interposer. We use

TABLE I: Dimensions of the 2.5D System

| Layers | Thickness | Materials | |
|---|---|---|---|
| Heat Sink | $6.9mm$ | | |
| Spreader | $1mm$ | | |
| Interface Material | $20\mu m$ | | |
| CMOS Chiplet Layer | $150\mu m$ | Silicon, Epoxy | |
| Microbump Layer | $10\mu m$ | Copper, Epoxy | |
| Silicon Interposer | $110\mu m$ | Silicon, Copper (TSV) | |
| C4 Layer | $70\mu m$ | Copper, Epoxy | |
| Organic Substrate | $200\mu m$ | FR-4 | |
| **Component** | **Diameter** | **Height** | **Pitch** |
| Microbumps | $25\mu m$ | $10\mu m$ | $50\mu m$ |
| TSVs | $10\mu m$ | $100\mu m$ | $50\mu m$ |
| C4 bumps | $250\mu m$ | $70\mu m$ | $600\mu m$ |

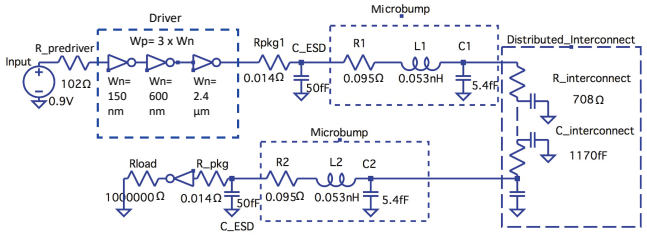


Fig. 2: *Inter-chiplet link model of 15mm length (based on a prior model [23]).*

DSENT [22] to calculate power of on-chip links and routers, and HSpice to compute power of inter-chiplet links based on a 2.5D interconnect model [23] (Fig. 2). We size up the drivers to ensure single-cycle propagation delay in the inter-chiplet links. The electrical mesh in the 2.5D system consumes up to $8.4W$, based on real benchmarks activities obtained from Sniper [24]. An electrical mesh network in a single-chip system with the same micro architecture consumes $3.9W$ (the low power values in both cases are due to low link activity). Essentially, here we trade off network power to match network performance in the 2.5D system with that in a single-chip system. This power increase, however, has negligible impact on the thermal profile of the whole system.

### B. Manufacturing Cost Model of 2.5D Systems

The cost benefit of 2.5D systems has already been discussed in prior work [9], [10], where a 20% to 30% reduction in cost can be achieved by replacing a single chip with a 4-chiplet 2.5D system. Smaller chiplets utilize more wafer area around the edge and achieve higher yield [9] leading to lower cost per unit area. Though one needs an interposer to integrate these small chiplets, the cost is rather low in case of a passive interposer (typically \$500 per $300mm$ diameter wafer [25]) because it can be manufactured using older process technologies [14], and with high yield (as much as 98% [26]).

Stow et al. [10] have proposed a cost model for 2.5D systems that takes into account the cost and yield of the CMOS chiplets, microbump bonding, and the interposer, assuming known good dies[1]. We adopt this model for estimating the cost of our 2.5D systems. Eqs. (1) through (4) [10] calculate CMOS dies per wafer and interposer dies per wafer, CMOS chiplet yield, CMOS per-chiplet cost and interposer cost, and the overall cost of a 2.5D system, respectively.

$$N_{CMOS} = \frac{\pi \times (\phi_{wafer}/2)^2}{A_{CMOS}} - \frac{\pi \times \phi_{wafer}}{\sqrt{2} \times A_{CMOS}}, \quad N_{int} = \frac{\pi \times (\phi_{wafer_{int}}/2)^2}{A_{int}} - \frac{\pi \times \phi_{wafer_{int}}}{\sqrt{2} \times A_{int}} \quad (1)$$

$$Y_{CMOS} = (1 + A_{CMOS} D_0/\alpha)^{-\alpha} \quad (2)$$

$$C_{CMOS} = C_{wafer}/N_{CMOS}/Y_{CMOS}, \quad C_{int} = C_{wafer_{int}}/N_{int}/Y_{int} \quad (3)$$

$$C_{2.5D} = \frac{C_{int} + \Sigma_{i=1}^{n}(C_{CMOS} + C_{bond})}{Y_{bond}^{n-1}} \quad (4)$$

[1]We do not explicitly model the testing cost. We assume that the testing costs of a single-chip system and a 2.5D system are similar because a 2.5D system costs less in per-chiplet testing but has an additional cost associated with testing the 2.5D system as a whole.

TABLE II: Notation used in Equations (1) through (10)

| Notation | Definition | Assumed Value |
|---|---|---|
| $\phi_{wafer}, \phi_{wafer_{int}}$ | Diameter of CMOS and interposer wafer | $300mm$ |
| $N_{CMOS}, N_{int}$ | CMOS and interposer dies per wafer | Eq. (1) |
| $D_0$ | Defect density | $0.25/mm^2$ [10] |
| $\alpha$ | Defect clustering parameter | 3 [10] |
| $Y_{int}$ | Yield of an interposer | 98% [26] |
| $Y_{CMOS}$ | Yield of a CMOS chiplet | from Eq. (2) |
| $C_{wafer}$ | CMOS wafer cost | $5000 [25] |
| $C_{wafer_{int}}$ | Interposer wafer cost | $500 [25] |
| $C_{int}, C_{CMOS}, C_{2D}$ | Chiplet, interposer, and single chip cost | from Eq. (3) |
| $Y_{bond}$ | Chiplet bonding yield | 99% [10] |
| $C_{2.5D}$ | Cost of the 2.5D system | from Eq. (4) |
| $l_g$ | Guard band along each interposer edge | $1mm$ |
| $w_{2D}, h_{2D}$ | Width and height of the baseline single chip | $18mm$ |
| $w_{int}, h_{int}$ | Width and height of the interposer (in $mm$) | from Eq. (9) |
| $w_c, h_c$ | Width and height of the chiplets (in $mm$) | from Eq. (8) |

| Notation | Definition |
|---|---|
| $A_{CMOS}, A_{int}$ | CMOS, interposer die area |
| $C_{bond}$ | Bonding cost of a chiplet [27] |
| $r$ | Number of chiplets in a row or column |
| $n$ | Number of chiplets $n = r \times r$, $n \in \{4, 16\}$ |
| $F$ | Frequency set $\{1000, 800, 533, 400, 320MHz\}$ |
| $V$ | Corresponding voltage set $\{0.9, 0.87, 0.71, 0.63, 0.63V\}$ |
| $f$ | Operating frequency $f \in F$ |
| $p$ | Active core count $p \in \{32, 64, 96, 128, 160, 192, 224, 256\}$ |
| $IPS_{2.5D}, IPS_{2D}$ | Instructions per second (IPS) of 2.5D system and 2D system |
| $s_1, s_2, s_3$ | Chiplet spacings (Fig. 4(a)). $s_1 = s_2 = 0$ for 4-chiplet case |
| $T_{peak}, T_{threshold}$ | Peak operating temperature and Temperature threshold for safety |

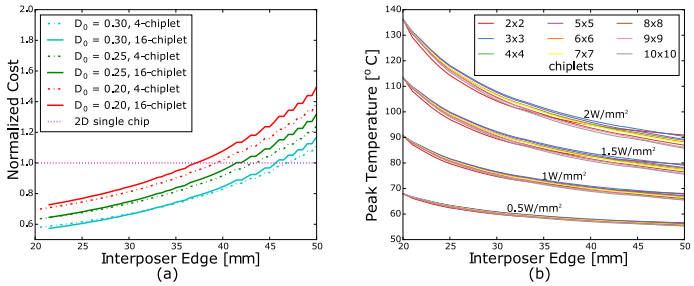


Fig. 3: *(a) Impact of defect densities on 2.5D system cost normalized to the single-chip system costs at the same defect densities. (b) Impact of chiplet counts, interposer sizes, and power densities on peak temperature of 2.5D systems with* ***uniform*** *spacing between chiplets.*

Fig. 3(a) shows the manufacturing cost of the 2.5D systems with various (square-shaped) interposer sizes normalized to an equivalent $18mm \times 18mm$ single-chip system for a range of defect densities [10]. The 2.5D system with a minimal interposer size has a cost saving ranging from 30% to 42%, compared to the cost of the single-chip system at the same defect density. The cost saving is higher for a larger defect density at which a single chip costs more due to its lower yield. Generally speaking, the 2.5D system cost increases as the interposer size increases.

### C. Thermal Behavior of 2.5D Systems

To understand the thermal behavior of a 2.5D system, we divide an $18mm \times 18mm$ single chip into $r \times r$ identical chiplets ($r$ varies from 2 to 10), place the chiplets onto an interposer in a matrix fashion with **uniform** spacing between adjacent chiplets, and leave $1mm$ guard band along each interposer edge. We vary the interposer edge length from $20mm$ to $50mm$ in steps of $1mm$ and calculate the corresponding spacing between chiplets. For a given interposer size, as the chiplet count increases, the spacing between the chiplets decreases. We assign synthetic power densities from $0.5W/mm^2$ to $2.0W/mm^2$ to the chiplets and perform thermal simulations (using HotSpot [28]) to get a better understanding of the thermal trends in 2.5D systems.

Fig. 3(b) shows the impact of chiplet counts, interposer sizes, and power densities on peak temperature of 2.5D systems. In general, for the same chiplet count and interposer size, the peak temperature increases with power density. For the same chiplet count and power density, as the interposer size increases, the peak temperature decreases due to the increased spacing between chiplets. For the same interposer size and power density, the peak temperature decreases with increasing chiplet count. It should be noted that in our 2.5D multi-chiplet system, the chiplets have high power density and the regions between chiplets do not dissipate power. Inserting spacing between chiplets helps with heat dissipation, but heat will still aggregate and form hotspots in the regions of high power density. Thus, we need to place the chiplets in a thermally-aware fashion.

Although a single chip with the *same power profile* and the *same area* as our 2.5D system would achieve a similar thermal profile, the single-chip solution is not the best choice from a cost perspective. For example, based on Eqs. (1)-(4) and parameters in Table II, increasing the single chip size from $20mm \times 20mm$ to $40mm \times 40mm$ results in $27\times$ higher cost because of drastically lower yield. Alternatively, an equivalent 2.5D system with four smaller chiplets and a $40mm \times 40mm$ passive silicon interposer has 27% lower cost (where the interposer cost is 30% of the 2.5D system) than a $20mm \times 20mm$ single chip.

From the cost perspective, as chiplet count increases in a 2.5D system, the time for the serial bonding process increases and the overall bonding yield drops, which increases the cost. Due to the limited thermal advantages of increasing chiplet count beyond $4 \times 4$ and the bonding yield consideration, in the rest of this work, we only consider 2.5D systems with 4 and 16 chiplets.

### D. Optimization of Chiplet Organization

To determine the optimal thermally-aware chiplet organization (including chiplet count, chiplet placement, active core count, and operating frequency), we formulate an objective function that maximizes system performance while minimizing system cost (see Eq. (5)). In Eq. (5), 2.5D system performance (in terms of instructions per second (IPS)) and cost are normalized to the baseline single-chip system, and the user-specified weight factors $\alpha$ and $\beta$ have no units. The objective function is subject to a variety of constraints listed in Eqs. (6) to (10).

$$Minimize: \quad \alpha \times \frac{IPS_{2D}}{IPS_{2.5D}(f,p)} + \beta \times \frac{C_{2.5D}(n,s_1,s_2,s_3)}{C_{2D}} \tag{5}$$

$$Subject\ to: \quad T_{peak}(f,p,n,s_1,s_2,s_3) <= T_{threshold} \tag{6}$$

$$w_{int} <= 50, \quad h_{int} <= 50 \tag{7}$$

$$w_c = \frac{w_{2D}}{r}, h_c = \frac{h_{2D}}{r} \tag{8}$$

$$w_{int} = w_c \times r + 2 \times s_1 + s_3 + 2 \times l_g, h_{int} = h_c \times r + 2 \times s_1 + s_3 + 2 \times l_g \tag{9}$$

$$2 \times s_1 + s_3 - 2 \times s_2 > 0 \tag{10}$$

Eq. (6) is the peak temperature constraint for a valid chiplet organization. Eq. (7) limits the interposer size to be no larger than $50mm \times 50mm$. This is within the exposure field size of 2X JetStep Wafer Stepper [29], which avoids extra stitching cost. We consider all chiplet organizations on an interposer that are axially and diagonally symmetric and we use Mintemp [20] workload allocation policy for our analysis, which minimizes operating temperature by assigning threads starting from outer rows or columns and then moving to inner rows or columns of the whole system in a chessboard manner. Eq. (8) calculates the chiplet width and height. Eq. (9) calculates the interposer width and height as a function of chiplet spacings ($s_1$, $s_2$, and $s_3$ in Fig. 4(a), which vary **independently**). Eq. (10) ensures there is no overlap between center chiplets. The 2.5D system cost is calculated using Eqs. (1) to (4).

**Pseudocode: Multi-Start Greedy Approach**

```
1) calculate cost and performance of 2.5D system for all (f, p, C_2.5D) combinations
2) input obj. func. weights (α,β)
   sort (f, p, C_2.5D) combinations based on obj. func. from low to high
3) foreach (f, p, C_2.5D) combination in the sorted order do
     generate random start points of (s1, s2, s3)
     foreach start point (S_current) do
       evaluate peak temperature T of S_current
       repeat
         generate a random neighbor placement (S_neighbor)
         evaluate peak temperature T' of S_neighbor
         if T' < T_threshold then
           output S_neighbor and (f, p, C_2.5D) combination and exit
         if T' < T then
           update minimum peak temperature T ← T'
           update current placement S_current ← S_neighbor
       until T < peak temperature of all the neighbor placements
     end for
   end for
```

To determine the minimum value of the objective function, an exhaustive search approach takes 180$k$ CPU hours (a calendar month with 250 computers running in parallel) to run thermal simulations for the whole design space of our example 256-core 2.5D system. The simulation time is long because there are over 680$k$ chiplet organizations (17$k$ chiplet placement options with 0.5$mm$ granularity, five voltage/frequency levels, and eight different active core counts) for each benchmark, and each organization takes up to 2 *mins* for a thermal simulation. Note that it takes 1.5$k$ CPU hours in total to simulate performance for all the 40 $(f,p)$ pairs using Sniper [24], an architecture-level simulator, when running the benchmarks listed in Sec. IV, which is insignificant compared to thermal simulation time.

To speed up the process of finding a solution to our optimization problem, we design a multi-start greedy approach to reduce the number of thermal simulations (see Pseudocode). Our approach has three steps. In the first step, we calculate the performance of the 256-core system for all 40 $(f,p)$ pairs using Sniper [24], and the cost ($C_{2.5D}$) of both 4-chiplet and 16-chiplet cases for discretized interposer sizes from 20$mm$ to 50$mm$ with 0.5$mm$ granularity using Eqs. (1) to (4). In the second step, we compute the objective function value for each $(f,p,C_{2.5D})$ combination using user-specified weights $\alpha$ and $\beta$ and sort these $(f,p,C_{2.5D})$ combinations in ascending order of objective function values. In the third step, we go through the list of $(f,p,C_{2.5D})$ combinations in the sorted order to find a chiplet organization that meets the temperature threshold. Here, for each $(f,p,C_{2.5D})$ combination, we use $m$ starting points (for each starting point, spacing values $s_1, s_2$ and $s_3$ are randomly picked), and we greedily explore the design space from these starting points. Each starting point ($S_{current}$) has six neighboring points (obtained by varying one of $s_1, s_2$ or $s_3$ by $\pm 0.5mm$). We randomly[2] pick one neighbor ($S_{neighbor}$) and evaluate the peak temperature of $S_{current}$ and this $S_{neighbor}$. If the neighbor has a peak temperature lower than the temperature threshold, it is a chiplet placement solution for the current $(f,p,C_{2.5D})$ combination. We then stop the process and pick this organization as our solution. If $S_{neighbor}$ has a lower peak temperature than $S_{current}$ (but higher than the temperature threshold), we make $S_{neighbor}$ the next $S_{current}$ and repeat the substeps mentioned earlier to check if it has a neighbor with lower peak temperature. If $S_{neighbor}$ has higher temperature than $S_{current}$, we pick another neighbor of $S_{current}$ and evaluate its peak temperature. If all neighbors of $S_{current}$ have higher peak temperature than $S_{current}$, we move on to the next random starting point. If there is no feasible solution among all the $m$ starting points, then we go to the next $(f,p,C_{2.5D})$ combination. If none of the $(f,p,C_{2.5D})$ combinations lead to a feasible solution, it means that the manycore system is unable to run at any $(f,p)$ pair within the given temperature threshold.

[2]We randomly pick the neighbor placement because out of the six neighbors, the neighbor that has the lowest peak temperature may not necessarily lead to a local minimum. We also avoid any biases resulting from evaluating neighbors in a fixed order.

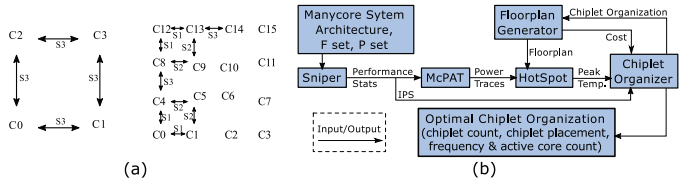


Fig. 4: *(a) Chiplet count and placement options. We vary the chiplet spacings* ***independently*** *to find the optimal chiplet placement. (b) Evaluation framework.*

We validate our multi-start greedy algorithm by comparing with the exhaustive search approach. The greedy algorithm with ten starting points (there is a tradeoff between accuracy and speed for different number of starting points) achieves the same result as the exhaustive search 99% of the time. Using the multi-start greedy approach, we can reduce the thermal simulation time from 180$k$ to 0.45$k$ CPU hours (400× speedup), and speed up the total simulation time (Sniper and Hotspot simulations) by 100× compared to the exhaustive search approach.

## IV. Evaluation Methodology

Our evaluation framework is shown in Fig. 4(b). We use Sniper [24] for performance evaluation. We use multi-threaded benchmarks from SPLASH-2 (`cholesky`, `lu.cont`) [30], PARSEC (`blackscholes`, `swaptions`, `streamcluster`, `canneal`) [31], HPCCG1 (`hpccg`) [32], and UHPC (`shock`) [33] suites that cover workloads of various performance and power profiles. We use different frequency/voltage levels and different numbers of active cores (see Table II) while evaluating the 256-core system. For each $(f,p)$ pair of each benchmark, we simulate 10 billion instructions in the parallel region or the full region of interest (ROI) if it finishes earlier. We collect performance statistics for each core every 1$ms$.

We use McPAT [34] to calculate power consumption of each core based on the performance stats from Sniper. We calibrate the McPAT output with the measured power dissipation data of Intel SCC [19], scaled to 22$nm$. We assume that the idle cores enter sleep mode and consume negligible power (close to 0$W$). As discussed in Sec. III-A, we calculate network power using DSENT and Hspice.

We use HotSpot-6.0 [28] for our thermal simulations, which can model layers (where each layer is composed of heterogeneous materials) in a 3D structure [35]. We model 2.5D systems based on industry prototypes [13], [14]. We generate detailed floorplan files specifying material properties of all blocks in each layer. We treat each core as a single block of heat source and use a $64 \times 64$ grid for thermal simulations. For the interface material, the spreader, and the heat sink, we use the default conventions in HotSpot, assuming spreader edge size is 2× interposer's edge, and heat sink edge size is 2× spreader edge. We adjust the convective resistance of heat sink to keep heat transfer coefficient constant. We set temperature threshold to 85$^oC$ and ambient temperature to 45$^oC$.

We implement a temperature-dependent leakage power model in our thermal simulations. We use a linear leakage model extracted from published power and temperature data of Intel 22$nm$ processors [20]. We assume 30% of power is leakage at 60$^oC$. We adjust the leakage power of each core based on its

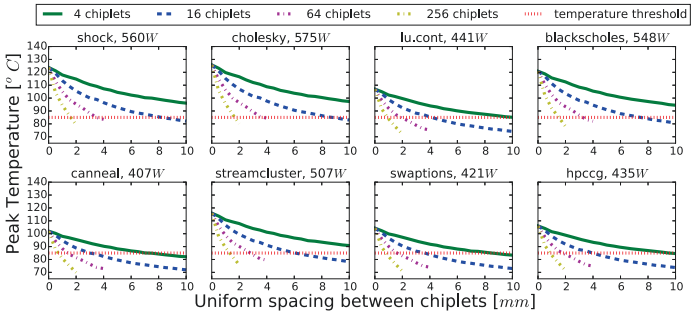


Fig. 5: *Peak temperature of a 256-core system with all cores active at 1GHz for single-chip case (0mm) and 2.5D integration cases for various chiplet counts and spacings (with chiplets placed in a matrix fashion).*

initial temperature obtained from HotSpot, and re-run HotSpot to update the thermal profile until the temperature converges.

In our evaluation framework (Fig. 4(b)), there is a closed loop between chiplet organizer, floorplan generator, and HotSpot. The chiplet organizer is implemented using the multi-start greedy algorithm (with ten starting points) as discussed in Sec. III-D. We use single-application workloads in this paper. In a more general case, a system runs a variety of applications. Our approach can be used by a designer to find the chiplet organization that minimizes the objective function for multi-application workloads. The designer could use the worst-case (i.e., the design with the largest interposer, which will ensure best performance for all applications), the average-case, or the weighted-average case (i.e., Eq. (5) becomes $\alpha \times \sum_i (\frac{IPS^i_{2D}}{IPS^i_{2.5D}} \times u_i) + \beta \times \frac{C_{2.5D}}{C_{2D}}$, where $i$ is the application index and $u_i$ indicates how frequently application $i$ runs) to select the optimal chiplet organization.

## V. Evaluation Results

### A. Peak Temperature Reduction using 2.5D Integration

We first study the impact of spacing between chiplets on the peak temperature (for different chiplet counts) with all cores active at $1GHz$ for various benchmarks (see Fig. 5). The $0mm$ spacing case refers to the single-chip system. For the 2.5D integration cases, we organize the chiplets in a matrix fashion with a **uniform** spacing (from $0.5mm$ to $10mm$ with a granularity of $0.5mm$) between adjacent chiplets, given the $50mm \times 50mm$ upper limit of the interposer size. As discussed in Sec. III, the 64-chiplet and 256-chiplet cases are not viable due to low overall bonding yield. We present them here to show the overall thermal trends.

The reported power values are the total power consumption under the single-chip case. These power values, which are unrealistic for 2D systems, can be viable for 2.5D systems from a thermal perspective: even at these large power consumption values, a 2.5D system can operate below a typical temperature threshold of $85^oC$. The challenge then will be the design of a power delivery network that can provide the current required for this large power consumption[3].

In general, for all 2.5D integration cases, the peak temperature decreases as chiplet spacing increases. High-power benchmarks need larger chiplet spacing to stay below the $85^oC$ threshold. For example, high-power benchmarks (`shock`, `blackscholes`, and `cholesky`) need a 16-chiplet system with $10mm$ spacing to meet the $85^oC$ constraint, while low-power benchmarks (`canneal` and `swaptions`) can easily meet the same constraint with 16 chiplets and $4mm$ spacing or with 4 chiplets and $8mm$ spacing. This analysis shows that even a naive chiplet organization can lower peak temperature significantly and provide opportunities to improve performance.

[3]Based on expert opinion [36], there are no fundamental limits in designing power delivery circuits for high-power chips (e.g., $500W$), but a number of engineering challenges would need to be addressed.

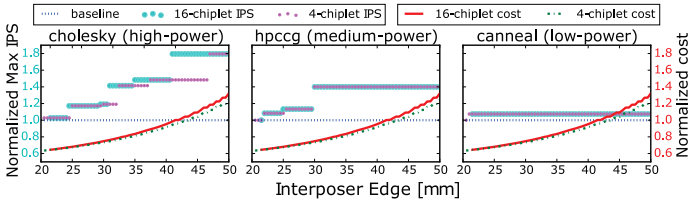


Fig. 6: *Maximum IPS and cost of 2.5D systems (normalized to maximum IPS and cost of a single-chip system) under $85^oC$ for example low-power, medium-power and high-power benchmarks. We only show results for 3 representative benchmarks due to space constraints. In total we evaluated 8 benchmarks.*

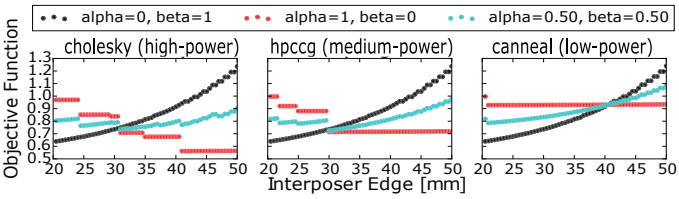


Fig. 7: *Minimum objective function (from Eq. (5)) value for different ($\alpha$,$\beta$) pairs across different interposer sizes for example low-power, medium-power and high-power benchmarks.*

### B. Balancing Performance and Cost of 2.5D Systems

In this subsection, we optimize the chiplet organization by considering **non-uniform** spacing between chiplets. Fig. 6 shows the normalized maximum IPS and cost of 2.5D systems. The maximum IPS, in general, remains unchanged as the interposer size increases, until the interposer size is large enough to find a chiplet placement that can operate the system at a higher performance level within the temperature threshold. The IPS curves have steps because we use discretized frequencies and active core counts. Since the cost of 2.5D systems only depends on the chiplet count and the size of interposer, the cost curves are the same across all benchmarks. With the minimum interposer size, the system cost decreases by 36% without performance loss. This reduction in cost is due to the higher yield of the smaller CMOS chiplets compared to the single-chip baseline.

At the same cost as the baseline, a thermally-aware 2.5D system with 16 chiplets can improve the performance by 41% on average across 8 benchmarks and by up to 87%. For the high-power benchmarks `shock`, `cholesky` and `blackscholes`, our approach achieves 87%, 80% and 75% performance improvement, respectively. As for the remaining benchmarks, our approach has 40% improvement for `Hpccg`, 24% for `swaptions`, and 14% for `streamcluster`; however, there is only 7% improvement for `canneal` and no performance gain for `lu.cont` when using 2.5D integration technology. The performance improvements for these benchmarks are limited because they do not need all cores active to maximize performance. For example, to achieve maximum performance, `canneal` needs 192 active cores, which is thermally feasible at small interposer sizes, while for `lu.cont` the maximum performance is achievable with 96 active cores even in conventional single-chip system under the temperature threshold. Although 2.5D systems do not bring performance benefits for `lu.cont`, our proposed thermally-aware chiplet organization can still provide lower operating temperature, which improves transistor lifetime and reliability.

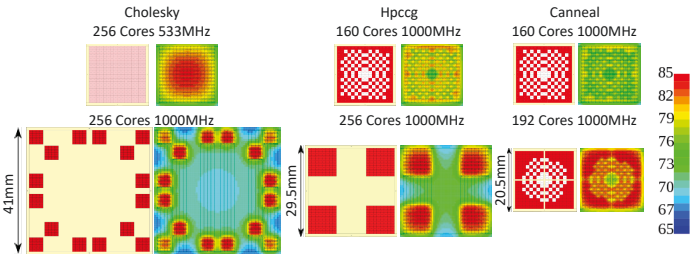


Fig. 8: *Choice of chiplet organizations that maximizes the performance under 85$^oC$ for single-chip baseline (top) and 2.5D systems (bottom).*

Fig. 7 shows the minimum objective function (Eq. (5)) values of three different choices for $\alpha$ and $\beta$ across different interposer sizes and different benchmarks. When $\alpha = 0$ and $\beta = 1$, the curves are the same as normalized minimum cost curves. When $\alpha = 1$ and $\beta = 0$, the curves are the same as inversed normalized maximum performance. When $\alpha = 0.5$ and $\beta = 0.5$, the objective function value is the weighted sum of $\frac{IPS_{2D}}{IPS_{2.5D}}$ and $\frac{C_{2.5D}}{C_{2D}}$. For a given pair of $\alpha$ and $\beta$, the optimal chiplet organization occurs at the minimum point on the objective function curve. For example, `cholesky` has the optimal organization at the interposer size of $31mm$, running at $1GHz$ with 192 active cores. The optimal chiplet organization, however, varies across benchmarks.

To choose the final chiplet organization, a designer would need to choose appropriate $\alpha$ and $\beta$ values. Fig. 8 shows examples of optimal chiplet organization and the workload allocation for $\alpha = 1$ and $\beta = 0$ under an 85$^o$C constraint. For `cholesky`, our technique improves performance by 80% by increasing frequency from $533MHz$ to $1GHz$, while the cost is similar compared to the baseline. For `hpccg`, our 2.5D system achieves 40% higher performance by increasing active core count from 160 to 256 and lowers cost by 28%. For `canneal`, the performance benefit is 7% because it saturates with 192 active cores; however, our approach reduces the cost by 36%. These results demonstrate that our thermally-aware chiplet organization technique can reclaim dark silicon by having more active cores and/or operate the cores at a higher frequency without violating the temperature threshold.

We analyze the sensitivity of our proposed approach to different temperature thresholds ranging from $75^oC$ to $105^oC$. The performance of the baseline single-chip system is lower at a lower temperature threshold, so there is more room for performance improvement. For the temperature thresholds of $75^oC$, $85^oC$, $95^oC$, and $105^oC$, our thermally-aware chiplet organization approach improves the performance by 41%, 41%, 27%, and 16%, respectively, on average across all 8 benchmarks.

## VI. Conclusion

Our work leverages thermally-aware chiplet organization in 2.5D systems to reclaim dark silicon. We propose to split a manycore system across multiple chiplets in the 2.5D system and then strategically insert spacing between the chiplets to reduce the operating temperature of the overall system, thus allowing more cores to operate at a higher frequency under the same safe peak temperature threshold. A multi-start greedy approach is used to determine the optimal chiplet organization that jointly maximizes performance and minimizes cost. Our analysis shows that for a 256-core system, compared to a single-chip design, our thermally-aware 2.5D integration approach improves performance by 41% (16%) on average and up to 87% (39%) without increasing the cost while staying below a peak temperature threshold of $85^oC$ ($105^oC$), or reduces system cost by 36%, without performance loss, at all temperature thresholds.

## Acknowledgments

This work was supported by NSF grants CCF-1716352, CCF-1564302, CCF-1149549, and CNS-1149703. We thank Prof. Ayse Coskun for many discussions and her help on thermal and design optimization aspects of the paper.

## References

[1] M. Shafique *et al.*, "The EDA challenges in the dark silicon era:temperature, reliability, and variability perspectives," in *Proc. DAC*, 2014, pp. 1–6.
[2] T. Muthukaruppan *et al.*, "Hierarchical power management for asymmetric multi-core in dark silicon era," in *Proc. DAC*, 2013, pp. 1–9.
[3] G. Venkatesh *et al.*, "Qscores: Trading dark silicon for scalable energy efficiency with quasi-specific cores," in *Proc. MICRO*, 2011, pp. 163–174.
[4] C. Silvano *et al.*, "Voltage island management in near threshold manycore architectures to mitigate dark silicon," in *Proc. DATE*, 2014, p. 201.
[5] J. Han *et al.*, "Approximate computing: An emerging paradigm for energy-efficient design," in *Proc. ETS*. IEEE, 2013, pp. 1–6.
[6] S. Pagani *et al.*, "TSP: thermal safe power: efficient power budgeting for many-core systems in dark silicon," in *Proc. CODES+ISSS*, 2014, p. 10.
[7] A. Raghavan *et al.*, "Computational sprinting," in *Proc. HPCA*, 2012, pp. 1–12.
[8] G. H. Loh, Y. Xie, and B. Black, "Processor design in 3D die-stacking technologies," *IEEE Micro*, vol. 27, no. 3, 2007.
[9] A. Kannan *et al.*, "Enabling interposer-based disintegration of multi-core processors," in *Proc. MICRO*, 2015, pp. 546–558.
[10] D. Stow *et al.*, "Cost analysis and cost-driven IP reuse methodology for SoC design based on 2.5D/3D integration," in *Proc. ICCAD*, 2016, p. 56.
[11] "DARPA CHIPS," http://www.darpa.mil/news-events/2016-07-19, 2016.
[12] P. Grani *et al.*, "Photonic interconnects for interposer-based 2.5D / 3D integrated systems on a chip," in *Proc. MEMSYS*, 2016, pp. 377–386.
[13] J. Charbonnier *et al.*, "High density 3D silicon interposer technology development and electrical characterization for high end applications," in *Proc. ESTC*, 2012, pp. 1–7.
[14] R. Chaware, K. Nagarajan, and S. Ramalingam, "Assembly and reliability challenges in 3D integration of 28nm FPGA die on a large high density 65nm passive interposer," in *Proc. ECTC*, 2012, pp. 279–283.
[15] W.-L. Hung *et al.*, "Thermal-aware floorplanning using genetic algorithms," in *Proc. ISQED*, 2005, pp. 634–639.
[16] M. Healy *et al.*, "Multiobjective microarchitectural floorplanning for 2D and 3D ICs," *IEEE TCAD*, vol. 26, no. 1, pp. 38–52, 2007.
[17] D. Cuesta *et al.*, "Thermal-aware floorplanning exploration for 3D multi-core architectures," in *Proc. GLSVLSI*, 2010, pp. 99–102.
[18] F. Frantz, L. Labrak, and I. O'Connor, "3D IC floorplanning: Automating optimization settings and exploring new thermal-aware management techniques," *Microelectronics Journal*, vol. 43, no. 6, pp. 423–432, 2012.
[19] J. Howard *et al.*, "A 48-core IA-32 processor in 45 nm CMOS using on-die message-passing and DVFS for performance and power scaling," *IEEE JSSC*, vol. 46, no. 1, pp. 173–183, 2011.
[20] T. Zhang *et al.*, "Thermal management of manycore systems with silicon-photonic networks," in *Proc. DATE*, 2014, pp. 1–6.
[21] Z. Zhang *et al.*, "Recent advances in flip-chip underfill: materials, process, and reliability," *IEEE Trans. Adv. Pack.*, vol. 27, no. 3, pp. 515–524, 2004.
[22] C. Sun *et al.*, "DSENT-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," in *Proc. NOCS*, 2012, pp. 201–210.
[23] M. A. Karim, P. D. Franzon, and A. Kumar, "Power comparison of 2D, 3D and 2.5D interconnect solutions and power optimization of interposer interconnects," in *Proc. ECTC*, 2013, pp. 860–866.
[24] T. E. Carlson *et al.*, "Sniper: exploring the level of abstraction for scalable and accurate parallel multi-core simulation," in *Proc. SC*, 2011, p. 52.
[25] G. Parès, "3D technology for photonics silicon interposer," in *Green IT workshop - Leti Days*, 2013.
[26] K. Tran *et al.*, "High-bandwidth memory white paper: Start your HBM/2.5D design today," Amkor Technology Inc., Tech. Rep., 2016.
[27] S. Farrens, "Wafer and die bonding technologies for 3d integration," in *Proc. MRS*, 2008, pp. 1112–E01.
[28] R. Zhang *et al.*, "Hotspot 6.0: Validation, acceleration and extension," 2015.
[29] K. Ruhmer *et al.*, "Lithography challenges for 2.5D interposer manufacturing," in *Proc. ECTC*, 2014, pp. 523–527.
[30] S. C. Woo *et al.*, "The SPLASH-2 programs: Characterization and methodological considerations," in *Proc. ISCA*, 1995, pp. 24–36.
[31] C. Bienia *et al.*, "The PARSEC benchmark suite: characterization and architectural implications," in *Proc. PACT*, 2008, pp. 72–81.
[32] M. Heroux, "HPCCG MicroApp," 2007.
[33] D. Campbell *et al.*, "Ubiquitous high performance computing: Challenge problems specification," GaTech, Tech. Rep. HR0011-10-C-0145, 2012.
[34] S. Li *et al.*, "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Proc. MICRO*, 2009, pp. 469–480.
[35] J. Meng, K. Kawakami, and A. K. Coskun, "Optimizing energy efficiency of 3-D multicore systems with stacked dram under power and thermal constraints," in *Proc. DAC*, 2012, pp. 648–655.
[36] E. Friedman, Personal Communication, Dec. 2016.