1. **Analyzing power-thermal-performance trade-offs in a high-performance 3D NoC architecture**

* The power consumption and thermal profiles of 3D NoCs can be improved by incorporating a Voltage Frequency Island (VFI)-based power management strategy and Reciprocal Design Symmetry (RDS)-based floor planning. In this paper, we undertake a detailed design space exploration for 3D NoC by considering power-thermalperformance (PTP) trade-offs.
* We specifically consider a small-world network-enabled 3D NoC (3D SWNoC) in this performance evaluation due to its superior performance and energy-efficiency compared to any other existing 3D NoC architectures.
* We demonstrate that the VFI-enabled 3D SWNoC lowers the energy-delay-product (EDP) by 57.3% on an average compared to a 2D MESH without VFI. Moreover, by incorporating VFI, we reduce the maximum temperature of 3D SWNoC by 15.2% on an average compared to the non-VFI counterpart. By complementing the VFI-based power management with RDS-based floor planning, the 3D SWNoC reduces the maximum temperature by 25.1% on an average compared to the non-VFI counterpart
* most of the existing VFI methodologies in 3D manycore chip use the conventional

MESH-based NoC architectures [7,8]. As shown later in this work,

incorporation of VFI-based power management for the MESH NoCs can

introduce significant performance penalty and energy overheads for

inter-VFI communications and thereby, fails to exploit the full benefits

offered by VFIs under a given performance constraint. Hence,

VFI-enabled 3D NoCs should incorporate a more efficient network topology

to minimize the performance penalty and maximize the potential

advantages of voltage-frequency (V/F) scaling.

* small-world network-enabled 3D NoC (3D SWNoC) is capable of

outperforming traditional MESH-based and other irregular counterparts. By adopting the power-law-based [11] small-world connectivity,

the SWNoC reduces the average hop count and communication path

length, and consequently, achieves significant performance gain and

lower energy consumption compared to traditional multi-hop MESH

NoCs.

* In a 3D SWNoC architecture with VFI-based power management,

the vertical links can act as the long-range shortcuts for inter-VFI data

exchange. This will ultimately mitigate the performance penalty inherent

in any VFI-based design. Moreover, by adopting the Reciprocal Design

Symmetry (RDS)-based floor planning [12] on top of VFI-based power

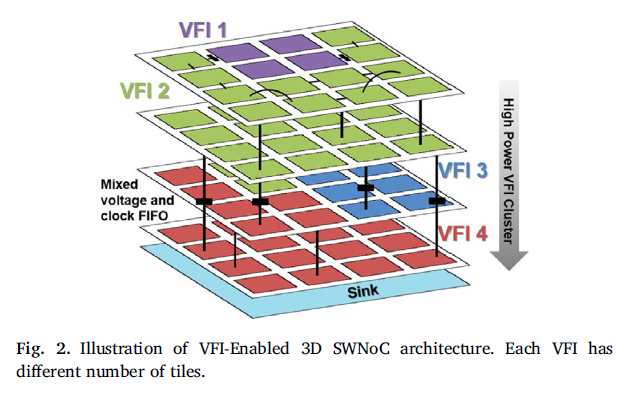
management, we improve the thermal profile even further.

* we incorporate the ALASH routing for SWNoC (Adaptive Layered Shortest Path Routing (ALASH))

and XYZ dimension order routing for MESH NoC.

**Voltage frequency island-enabled 3D NoC**

* A **Voltage Frequency Island (VFI)**-based power management strategy is used to reduce power consumption in systems-on-chip (SoCs) or multi-core processors by grouping cores or components into **voltage-frequency islands**. Each island operates at an independent voltage and frequency, optimized for its workload. This allows components with lower performance demands to run at reduced power, while higher-demand components can operate at full speed. By controlling power consumption at the island level, the strategy enhances energy efficiency without significantly affecting performance, particularly in heterogeneous systems.
* For example, cores with low utilization are grouped together and assigned low V/F level, while cores with high utilization are clustered together to allocate high V/F levels.



**Reciprocal** Design **Symmetry (RDS) placement**

* Stacking high power consuming cores directly on

top of each other increases the power density and consequently, creates

thermal hotspots. To solve this thermal issue in the 3D NoC, we employ

Reciprocal Design Symmetry (RDS)-based vertical placement. Fig. 3(a)

shows the general vertical placement in a 3D architecture where cores in

every die are stacked on top of each other (their individual sizes may vary

depending on the design). Hence, we call this general vertical placement

as the ‘STACK’ arrangement. In STACK placement, adjacent routers are

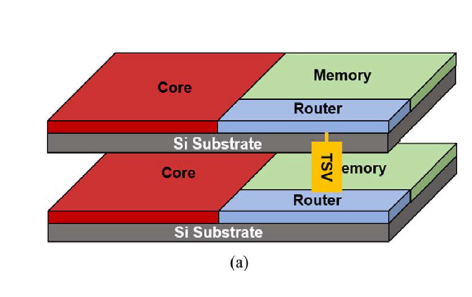
perfectly aligned and TSVs maintain direct communications among

them. Consequently, the routers from adjacent layers (dies) communicate

in a single hop with STACK-based placement. This ensures maximum

achievable performance from floor planning and router placement.

However, it may introduce thermal hotspots in the system.



* To alleviate the thermal hotspots due to stacking of high power

consuming cores on top of each other, RDS-based floor planning has been

advocated [12]. Fig. 3(b) shows the RDS-based placement in the 3D NoC

architecture. For RDS-based floor planning, instead of stacking cores on

top of each other (STACK), the cores are placed diagonally such that the

amount of overlap area among the vertically adjacent cores is minimized.

In general, cores dissipate the highest amount of power in a manycore

chip. Consequently, the diagonal vertical placement of cores and associated

routers alleviates creation of the thermal hotspots without any

additional area overhead. However, it is necessary to add an extra planar

link between two diagonally placed routers. As a result, the total interconnect

length increases and the overall performance worsens compared

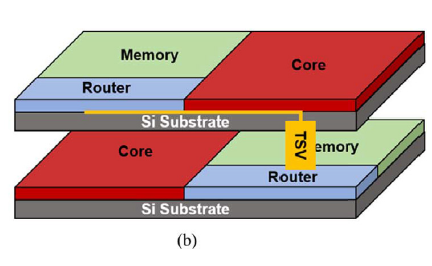
to the STACK-based design. Hence, we achieve thermal efficiency in the

RDS-based design at the cost of performance degradation. In this work,

we employ RDS vertical placement in 3D NoC architectures on top of the

VFI-based power management and evaluate the efficiency of the proposed

design considering the performance-thermal trade-offs.



**Thermal-aware network optimization**

we define a single cost function that combines them as follows:

cost = α⋅C + β⋅T (2)

where C is the unified NoC performance from (1); T is the temperature of

the system determined by following [34]; and α and β are weighting

factors. Our goal is to minimize the cost function to achieve the optimum

network latency, energy consumption, and the lowest temperature of the

3D SWNoC.

To optimize the overall cost function, we use simulated annealing

(SA) algorithm. As we focus on both core and link perturbation in the 3D

SWNoC architectural space, we utilize two moves for solution perturbation

namely the swapping (S(i,j)) and the new link placement (L(l)) as

discussed below:

\_ Swapping - S(i,j): selects two cores i and j randomly and swaps their

locations.

\_ New link placement - L(l): selects and removes a link of length-l between

two randomly selected routers. If a link is a planar link, create a

new link of the same length between two other routers in the same

layer. If a link is a TSV along the vertical direction, create a new TSV.

To incorporate the VFI-based power management, we modify the simulated annealing algorithm used for NoC architecture optimization.

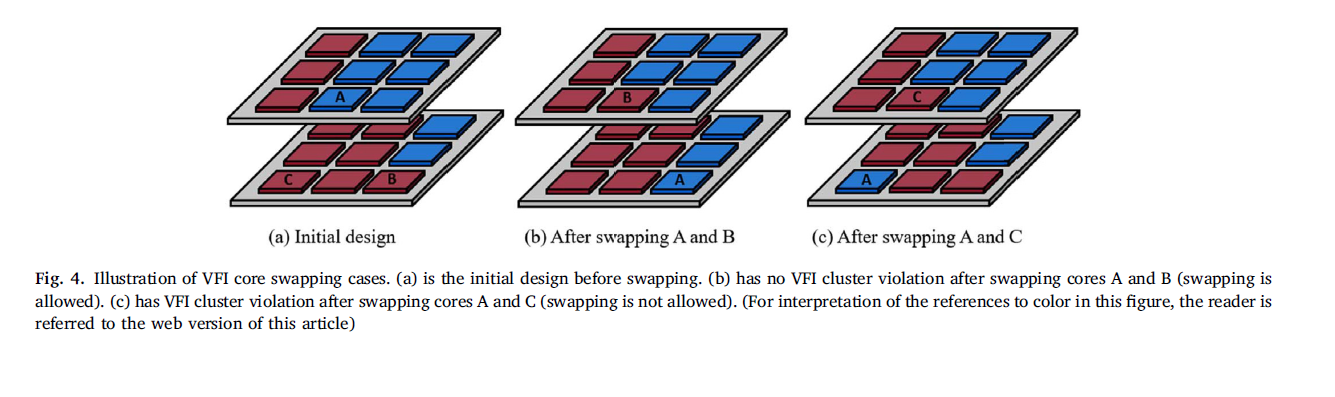
First, we modify the swapping perturbation function, S(i,j) to S0(i,j) as

follows:

\_ VFI Swapping – S0(i,j): selects two cores i and j randomly and swaps

their locations if the locations do not violate the constraint of VFI

clusters after swapping.



**Experimental results and analysis**

* For this performance evaluation, we consider two relevant performance metrics: latency, and

energy-delay-product (EDP). We define the EDP as the product of network latency and energy dissipation to unify the effects of both thesetwo metrics into a single parameter. To evaluate the thermal profiles of the architecture, we consider the maximum, average, and minimum temperature of the manycore chip.

* For ease of referencing, the joint optimized (performance and thermal) SWNoC and the baseline SWNoC (optimized only for performance) are marked as PTO and PO respectively.

**Effects of performance and thermal joint-optimization**

* **non-VFI system :** From the figure, we can see that the PTO-based design improves the

temperature profile compared to PO across all the benchmarks. The peak

temperature of PTO is lower than PO by 1.9–18.9% (2.3 C to 19.8 C)

depending on the specific benchmark. In addition, the difference between

the maximum and minimum temperatures of PTO is lower than

that of PO. The PTO-based NoC more efficiently distributes the total

power over the whole 3D chip when compared to the NoC optimized for

performance (PO) only, and hence, PTO alleviates creation of thermal

hotspots.

* In addition, to compare the performance of PO and PTO, Fig. 7 shows

their respective EDP values. For comparison purpose, the EDPs are

normalized with respect to the EDP value of PO. Here, to compute the

EDP value, we consider the average message latency and the average

message energy. As shown in the figure, the PTO shows on an average

2.2% higher EDP values than that of the PO, while the maximum increase

of 4.4% is observed for the CANNEAL benchmark.

By comparing both Figs. 6 and 7, one can observe that the PTO

technique significantly improves the thermal efficiency (maximum reduction of peak temperature by 18.9%) of the 3D SWNoC while only

incurring a maximum 4.4% EDP penalty when compared to the baseline

PO. In addition, if we compare the average of the peak temperature and

EDP values, then PTO shows 9.9% thermal improvement while incurring

2.2% higher EDP compared to PO. Hence, we can conclude that PTObased

design methodology is more efficient considering the

performance-thermal trade-offs compared to the PO-based approach.

**VFI-based power management.**

* we can observe that the EDP value also decreases for

both the MESH- and SW-based architectures as the number of layers increases.

The EDP value of a network depends on the average hop count

and communication path length. As seen from Fig. 9, as the number of

layers increases, more number of vertical links can act as the long-range

shortcuts between layers and the path length reduces. Consequently,

following the network latency trend in Fig. 8, the EDP also decreases

progressively as the number of the layer increases. The NoC with more

layers offers more connectivity and opportunities for exploring the

optimized NoC architecture. Hence, 3D NoCs with higher number of

layers perform better than 2D NoCs (number of layer equals to 1 for 2D

NoC).

* moreover, as shown in Fig. 10, the EDP of the VFI-enabled SWNoC is

lower than that of the VFI-enabled MESH by up to 30.8% (average

16.9%) depending on the specific benchmark for the same number of

layers. The lower EDP of SWNoC is primarily due to its lower path length

compared to the MESH (can be observed in Fig. 9). Lower path lengths

lead to lower inter-router hop counts resulting in the reduction of energy

consumption in the routers and links. Thus, the SWNoC always shows

lower EDP than the MESH. Also, the latency penalty incurred by the VFI

technique is minimized by adopting the SW-network topology.

* From the figure, it is evident that

among all these configurations, VFI SWNoC with 4-layer achieves the

lowest EDP. On an average, VFI SWNoC achieves 57.3% lower EDP

compared to NVFI 2D MESH (with 1-layer).

**Evaluation of thermal profiles**

To demonstrate the merits of VFIbased

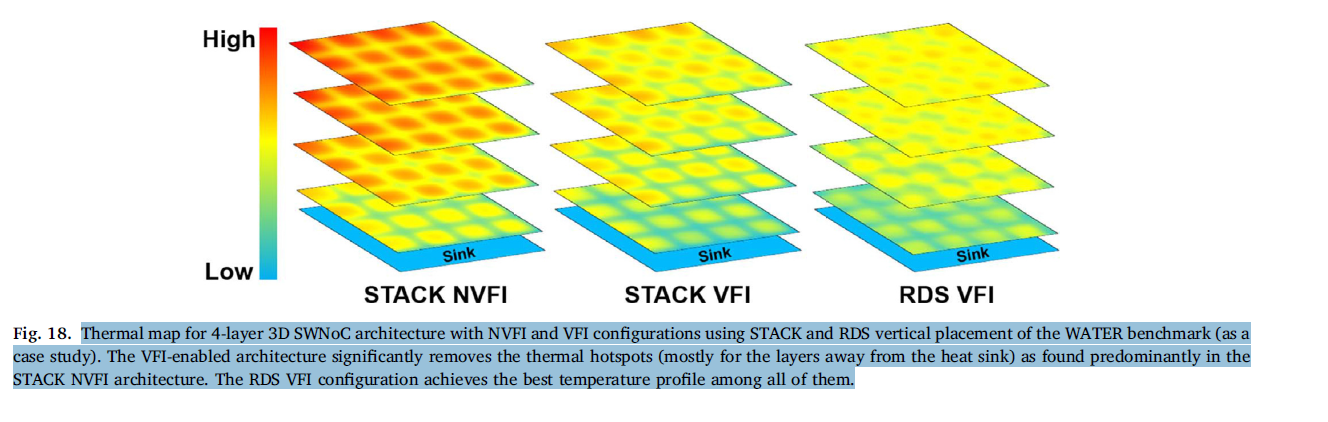
power management to improve the thermal profiles, we consider

the 3D SWNoC architecture with NVFI and VFI configurations. In addition,

to evaluate the joint effects of power management and vertical

placement, we compare the thermal profiles of VFI configuration using

STACK and RDS-based floor planning



* As the number of layers increases, the

possibility of thermal hotspot creation increases for STACK configurations.

However, for RDS-based floor planning, the creations of hotspots

are avoided. Consequently, the average and maximum temperatures are

reduced for the RDS-based configurations.

* As discussed in Section 5.3.1, VFI-based power management significantly

reduces overall power consumption of the manycore chip and

consequently, the temperature profile of the system improves. In addition

to the VFI-based methodology, RDS-based vertical placement also helps

to reduce the creation of thermal hotspots

* By combining the result from Figs. 14 and 15, it is

seen that the SWNoC with VFI configuration using RDS placement shows

on an average 25.1% less maximum temperature than that of the STACK

NVFI configuration.

**Trade-off summary**

* We demonstrated that, the VFI-enabled power management

methodology along with the RDS-based floor planning achieves the

best PTP trade-offs for the 3D manycore chip.

* We demonstrate that the VFI-enabled 3D SWNoC achieves on an average

57.3% lower EDP than conventional 2D MESH-based architecture. By

incorporating RDS-based floor planning, we alleviate the thermal hotspots

and thereby, improve the overall thermal profiles. The 3D SWNoC

architecture while incorporating both the VFI and RDS techniques lowers

the maximum temperature by 25.1% on an average compared to the

conventional non-VFI counterpart.

* Finally, the proposed performancethermal

co-design (PTO) approach for 3D SWNoC improves the thermal

profile by 18.9% while incurring only 2.2% performance penalty

compared to only performance oriented (PO) 3D SWNoC architecture.

**Limitations**

* **Focus on Specific Architectures**: The study primarily evaluates the small-world network-enabled 3D NoC (3D SWNoC) and its performance compared to 2D MESH architectures. This narrow focus may limit the generalizability of the findings to other types of NoC architectures, which might behave differently under similar conditions