
Article:
Crispin-Bailey, Christopher orcid.org/0000-0003-0613-9698, Thuphairo, Pakon, Austin, Jim orcid.org/0000-0001-5762-8614 et al. (2 more authors) (2025) Exploring the Feasibility of Novel 3D Polyhedrally-Tiled Computing Arrays. IEEE Access. 11146660. pp. 159685-159713. ISSN: 2169-3536

https://doi.org/10.1109/ACCESS.2025.3605303

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 June 2025, accepted 31 July 2025, date of publication 2 September 2025, date of current version 17 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3605303

PERSPECTIVE

Exploring the Feasibility of Novel 3D Polyhedrally-Tiled Computing Arrays

CHRIS CRISPIN-BAILEY[1], PAKON THUPHAIRO[2], STEVEN WRIGHT[1], ANTHONY MOULDS[1], AND JIM AUSTIN[3]

[1]Department of Computer Science, University of York, YO10 5DD York, U.K.
[2]Department of Computer Engineering, Faculty of Engineering, Rajamangala University of Technology Rattanakosin, Nakhon Pathom 73170, Thailand
[3]Department of Computer Science, University of York, YO10 5DD York, U.K. *(Retired)*

Corresponding author: Chris Crispin-Bailey (chrisb@cs.york.ac.uk)

This work was supported in part by the Royal Thai Government National Science and Technology Development Agency (NSTDA) Ph.D. Scholarship of Office of Educational Affairs (OEA) under Grant ST_G5599, and in part by U.K. Government through Engineering and Phsyical Sciences Research Council (EPSRC) under Grant G0103201 IFPT-2023-04. The work of Pakon Thuphairo was supported by the Faculty of Engineering, Rajamangala University of Technology Rattanakosin.

ABSTRACT Polyhedral tiled computing arrays (PTCAs) are a largely unexplored paradigm in which the assembly of high performance computing (HPC) structures may be achieved by the multidimensional tiling of particular polyhedral modules, each housing computing devices. Typically, PTCAs form *physical* three-dimensional (3D) topologies which readily support *logical* three-dimensional topologies. PTCAs eliminate the need for traditional backplanes, racks and custom host circuit-board modules, while inherently composing coherent and scalable input/output (IO), power and cooling grids simply by abutting to neighboring modules in one or more dimensions. This highly novel concept offers unique possibilities in all three of those domains, and therefore quantifying PTCA capabilities and limitations is critical in establishing the desirability of such systems for future modular and heterogeneous HPC systems. The questions must then arise: *is this a realistic paradigm?*, and, *can such a system be practically engineered?* In this paper we contribute a number of insights, an analytical methodology for the evaluation of a specific class of PTCAs, based upon simple-cube and truncated octahedral modules arranged into cubic arrays, and we demonstrate the use of numerical methods and formulae to analyze the potential capabilities and limits of such novel systems with both existing reference points and future technology expectations. Whilst mapping out all aspects of this novel design-space is somewhat speculative, and beyond the scope of a single paper, we do conclude by identifying key challenges and 'road-map' goals aimed toward finally reaching that objective.

INDEX TERMS 3D mesh arrays, future computing, HPC, interconnection network, network topology, polyhedra, power density, system cooling, three-dimensional interconnect.

I. INTRODUCTION

Traditionally, the most convenient route to construct a large-scale computer system has been to use a cabinet, back-plane, rack, and circuit board assembly: either from highly standardized components or via custom rack-based printed circuit board (PCB) designs, engineered for each specific solution. Although mainstream, this approach introduces multiple levels of technical and engineering hierarchy even before assembly can be considered complete as a system, and effectively imposes 2D models of assembly onto systems that are often logically structured as 3D topologies in actuality.

Given the direction of Moore's law, Dennard scaling, VLSI technology road-maps, and Koomey's law, it is now timely to better understand new paradigms and potentially identify radically new concepts for future high-performance computing (HPC) systems. One such possibility relies upon polyhedral tiles: 3D forms that themselves tile into 3D structures. There is, at best, sparse literature indicating work in this field to date, indicating novelty and a need for innovation and exploration of this design space to establish its feasibility.

The associate editor coordinating the review of this manuscript and approving it for publication was Tomas F. Pena.

Previous work undertaken in this area shows the potential for large-scale 3D system assemblies based upon polyhedral tiled computing array (PTCA) concepts, including power grid simulations, evaluations of wired and wireless connectivity strategies, hardware prototypes and network/workload modeling comparisons to more traditional physical topologies [1], [2], [3], [4], [5]. However, there is now a need for a well-defined numerical approach to evaluating such systems on a theoretical basis, in order to understand their viability as genuine alternatives. This includes modeling of performance-relevant metrics such as the number of processors, IO bandwidths, power consumption, and FLOPS, but in the case of PTCAs it is also important to understand size, volume, power delivery, heat management and their related engineering challenges. This paper seeks to evaluate four key questions that are critical for arriving at useful conclusions about the viability of PTCAs in real-world deployment:

1) What are feasible shapes and sizes for cores?
2) Can cascaded power grids achieve required capacities?
3) Can tiled modules interface to each other effectively?
4) Is power density and cooling viable in such systems?

These are important questions, and can only be answered by either building non-trivial prototype systems, or by making use of analytical models and simulations to examine the detail of these parameters. Ultimately bringing both together would provide the ideal outcome.

Specifically, this paper makes the following contributions:

- A set of mathematical equations is presented that can be used to derive a number of useful properties of a PTCA;
- It explores possible PTCA configurations, evaluating IO, power, and physical parameters, and constraints for a range of PTCA array sizes;
- A set of use-case evaluations is detailed, showing specific design goals being translated into a PTCA paradigm and reporting on their system characteristics;
- An evaluation of power density versus size and core power is given, and implications for current and future petascale and exascale PTCA systems are considered.
- Finally, a number of 'road-map' recommendations are given for more detailed studies into specific aspects of PTCA design space, suggesting a set of 'road-map goals' aimed toward a complete understanding of this novel paradigm.

The remainder of this paper is structured as follows: Section II outlines the background and motivations for PTCA; Section III introduces a series of mathematical equations for predicting the physical and computational characteristics of PTCAs; Section IV demonstrates the use of these mathematical models to evaluate use-cases inspired by relevant and well understood real world HPC systems; Section IV-E evaluates the potential for technological trends to shift the perspective over the coming 10-year horizon; Section V provides a summary of the necessary road-map objectives needed to move forward toward understanding and engineering of practical PTCA systems at large scales and, finally, Section VI concludes this paper with a final view of work presented and its implications.

A. SUMMARY OF KEY TERMS INTRODUCED

For convenience, the following quick introduction to key terms introduced in this paper may be useful to readers.

- **PTCA**: Polyhedral Tiled Computing Array. An array of polyhedral shaped/packaged computational modules.
- **Polyhedra/Polyhedron**: a 3D structural shape with geometric properties.
- **T-facet**: a trapezoidal (typically square) facet on the surface of a relevant polyhedron.
- **H-facet**: a hexagonal facet on the surface of a relevant polyhedron.
- **Single-packed array**: in this paper this refers to a 3D array where only T-facets are abutted in the tiling model, resulting in an incomplete space-packing outcome.
- **Double-packed array**: in this paper this refers to a 3D array where spaces between tiled polyhedra are filled with other polyhedra (involving both T and H-facet abutment), and achieving complete or high-density space packing outcomes.
- **Kelvin-Core / K-core**: a short-hand name given in this paper to a truncated octahedral polyhedron - a shape which has six 'T' and eight 'H' facets.
- **PCB**: Printed Circuit Board.
- **MCM**: Multi-Chip Module.
- **SIP**: System-In-Package.
- **SOC**: System on Chip.
- **FLOPS** Floating point operations per second.

II. BACKGROUND CONCEPTS AND MOTIVATION

A. MOTIVATION

Before expanding into detail, it is perhaps useful to highlight some key comparisons in terms of the approach which PTCA attempts to offer, as differentiated from other mainstream HPC examples. Consider Table 1, which compares a number of mainstream HPC cases against a generic PTCA model:

- Many existing systems depend upon blade or custom board level designs which aggregate multiple processing elements into local groups, and then associate them via rack-mount structures. Those modules are typically powered via a high-current backplane power bus.
- PTCA utilizes freely tileable modules, whereby abutting a number of modules forms a typically 3D physical array of varied possible overall structure (a 3D mesh being the most obvious).
- Power distribution in PTCA systems is via a collective 3D-scalable power grid, formed by tiling of modules.
- In mainstream systems, IO is often hierarchical - combining board-level groupings of processing elements with cable-based inter-modular connectivity, resulting in large numbers of cables with lengths of the order of metric metre or sub-metre scales.

- PTCA uses point-to-point centimetre-scale connection paths in a typically uniform grid structure.
- Mainstream cooling generally relies upon air-cooled cabinet designs, but increasingly the use of external conduits and 'bolt-on' cold-plates and fluid-delivery networks is commonplace. Full liquid-immersion is a newer approach in terms of mature large-scale solutions but growing in frequency of deployment.
- PTCA envisages internalized cooling voids passing through the inner space within the tiled module, and the formation of contiguous air/liquid cooling paths by virtue of modules abutting into array structures.

From these basic observations, one can see that PTCA sits in a very different corner of the design space than traditional and mainstream HPC solutions, with multiple differentiating factors. The following sections will expand these comparisons in significantly more detail.

Traditional rack-mount systems are widely used and have the convenience of a well-tested mode of assembly, but also come with significant limitations. A primary issue is that processing and auxiliary components must be aggregated via rack-mount PCB planes, themselves complex design endeavors, and any such host-board design must also be constrained by rack and backplane requirements rather than being able to directly mirror the intended logical topology of the system. This hinders the literal implementation of logically three-dimensional (3D) computing topologies, obliging them to map onto what are inherently closer to two-dimensional (2D) signaling structures such as rack and backplane, and/or heavily augmented by complex inter-module IO cabling. This often leads to design compromises that are undesired and potentially detrimental to performance and cost.

Such constraints are tolerated in lieu of the convenience of having a well-established standard construction medium and accepting the compromises resulting from mapping logically 3D topologies onto '2D' structures. Some systems aggregate such data flows into common data-channels via the backplane or inter-board and inter-cabinet data highways. Likewise, these are at best described as '2.5D' systems assemblies in comparison to their logical 3D topologies. The SpiNNaker and MDGRAPE systems are examples where this approach is observable [6], [9]. On the other hand, relative ease of maintainability in a rack system is a positive aspect that should also be noted.

Meanwhile, more loosely-coupled systems may route inter-node messaging onto traffic-aggregating switching networks such as fat-tree routing structures, but cannot facilitate direct connectivity between cores in 3D physical space. Consequently, extra costs are introduced in terms of message structure, buffering, latency, hardware component expense, power consumption, heat, physical space, and other factors. With the evolution of ever more complex AI and big-data systems these issues will only become more salient in the coming decade.

In strong contrast, PTCAs are defined as true 3D assemblies, and permit direct point-to-point connection to neighboring nodes in all three dimensions at large scales with uniformity. This is achieved by employing a directly three-dimensional modular system assembly paradigm founded upon the principle of composing arrays of modular polyhedral computing nodes from within the class of tileable polyhedra.

The nature of such modules allows direct connection to nearest neighbors in three dimensions for power, IO and system cooling. All three of these requirements may then aggregate into their respective networks simply by abutment of such modules. Compositions of system-level PCB planes are not required, and rack-mount and backplane concepts are eliminated in any traditional sense. Of course there is a trade-off, in that PTCA systems lose some of the generality of switch-routed topological implementations, and maintainability must be managed in different ways, but there are nonetheless valuable gains where tightly-coupled grid-like fabrics are desirable.

While modules that are simple cubes might seem the obvious choice of modular shape for such a paradigm, and having the advantage of plain simplicity, there are other partial or fully space-filling polyhedra, some having distinctive capabilities including extra-dimensional properties beyond that of a cube-based X-Y-Z grid structure.

Tiled modules therefore offer the possibility of overcoming such serious limitations, simultaneously solving three key problems – logical versus physical IO mapping, power distribution, and heat dispersion, simply by adopting a polyhedron-based assembly model.

B. RELATED WORK

Significant insights and motivations may be found in the work of earlier studies, highlighting the key advantages and necessities of moving scaling into higher dimensions to meet the needs of increasingly complex future systems, with an important observation made by those authors:

"We find that historical efficiency trends are related to density and that current transistors are small enough for zetascale systems once communication and supply networks are simultaneously optimised. We infer that biological efficiencies for information processing can be reached by 2060 with ultra-compact space-filled systems that make use of brain-inspired packaging and allometric scaling laws". [13] (published in 2011).

Meanwhile, a recent white paper on future unconventional HPC systems by leading researchers, also identifies some important key requirements to meet the needs of future HPC systems over the next decade [14]. These include, in particular:

- Short low-latency electrical links
- Co-packaged optical interconnects
- Modular fabrics
- Disruptive heterogeneity
- Flexible wire-composable systems

TABLE 1. Comparison of some key design factors for mainstream and PTCA systems.

MODEL	Compute Node IO Hierarchy	Power Delivery	Cooling
Spinnaker [6]	Inter-Nodal PCB Inter Board Custom Backbone Inter-Rack Cable IO paths up to 0.5m scale	Rack and cabinet Power Bus	Air cooled Cabinet
Google TPU Data Centre [7], [8]	Cable, Direct Optical beam interchange	Rack and cabinet Power Bus	Cold-Plate Liquid Cooling+air cooling
MDGRAPE-4A [9], [10]	Inter-CPU wired, Inter-board optical IO, IO paths 1 metre scale (est)	Rack and cabinet Power Bus	air-cooled
Anton-3 [11], [12]	Inter-board High Speed IO cabling, IO paths 1 metre scale (est)	Rack and cabinet Power Bus	Liquid Cooling, Cold plate
PTCA	Infra-nodal PCB, Point-to-Point Dedicated Links, IO paths on 100 mm scale	Inter-nodal distributed grid	Air or liquid internalized cooling grid, direct immersion cooling

The authors argue that, in the context of both of these perspectives, co-integration of cooling solutions are also of high priority, and would add this to the list as a further important challenge to be met, especially given the diverse range of technologies and the granularity of solutions at play in that domain [15].

It is useful to make some observations upon the limitations of rack-mount assembly, which are easily observed in some successful, well known and relatively large-scale systems, which adopt fairly standard rack and backplane approaches. For example, in the SpiNNaker project [6], [16], [17], [18], [19], the SpiNNaker 'million core' system is built from a number of adjacent and linked server cabinets, with each cabinet containing five rackmount PCB clusters. Each rackmount cabinet contains an array of identical SpiNNaker multicore SOCs (system-on-chip) arranged as twenty four 48-chip PCBs in racks and cabinets with five modules per cabinet. In computational terms however, SpiNNaker is inherently a 3D logical topology (a hexagonal torus), and achieves connectivity between SOCs on different rack modules via multiplexed (SpiNNlink) backplane routed data highways, and also by connection between the server cabinets, reportedly requiring 'thousands of metres' of cabling [17]. A considerable amount of hierarchical design and infrastructure is thus imposed upon the implementation, which ultimately aims to combine many processing nodes into a collective system array.

Cooling is an equally complex problem in large-scale array designs, with hierarchical and inter-modular issues analogous to that of data connectivity. For example, the Google TPU server has cooling pipes that must connect across multiple domains in order to form a large-scale cooling network [8]. This cold-plate directed liquid cooling approach is essential given the power density of the Google TPU units. These approaches seek to overcome the power density limitations of air cooling, where server cabinets can struggle to achieve power densities beyond the region of 40 kW unless employing external air-chiller units or internal highly directed cooling approaches. Fluid-assisted cold-plate cooling may push limits as high as 100 kWh [20], [21], while bulk immersion systems can achieve power densities approaching 200 kW for volumes similar to a 42U rackmount cabinet ($\approx 1m^3$). When exploring new and novel design concepts, it is anticipated that cooling strategies will require detailed modeling. Examples of such work for more traditional architectures include modeling and optimizing workload (and thus power) distribution within arrays, and detailed modeling at the component and board level [22], [23], [24]. The need for flow dynamics (be it air or liquid) is an important requirement for 3D grid concepts, especially as dimensions scale downward and effects such as turbulence and flow-resistance within cooling voids become more significant relative to general liquid or air flow models.

Another notable system in this domain, MDGRAPE-4A, occupies a different end of the spectrum to SpiNNaker (which uses devices with power consumption of the order of only a few Watts), instead relying upon a much smaller array of albeit far more powerful processing nodes, consuming of the order of 120 W (with core SOC power of 85 W), in contrast to the reported 1 W for the SpiNNaker SOC [9], [19]. For comparison, the latest Google TPU v4 devices are approaching 200 W power consumption [7]. Likewise, the Anton series of HPC systems also typically scale up to 512 nodes with a similar scale and complexity to MDGRAPE-4A, i.e., fewer but more powerful nodes, again arranged in a 3D logical structure, but using highly optimized custom ASIC components in place of more general commodity CPUs [10], [11], [12]. The Google TPUv4 installation also operates on a grid-wise principle, using a 3D torus, though using an advanced optical switching network to link at a higher level of hierarchy [7]. All of these aforementioned and very successful systems are arranged in PCB planes, racks, cabinets, and inter-cabinet connectivities, with similar 2D physical trade-offs versus a true 3D logical topology.

More generally, 3D logical topologies, including 3D mesh variants, have been of great interest in a wide range of

computing problems, including scientific modeling problems, big data, image processing, and increasingly complex AI systems. These systems vary widely in scale: SpiNNaker, for example, achieves an impressive 'million core' system scale, and yet is orders of magnitude smaller in scale than future neuromorphic systems might require, while MDGRAPE-4A, BlueGene, Google TPU, and others use hundreds of more powerful processors [6], [7], [9], [10], [11], [25], [26], [27], [28], [29].

Many computational application domains have 3D *logical* workload topologies, including volumetric medical imaging, molecular modeling, drug discovery, protein modeling, climate modeling, aerodynamics and fluid dynamics, among others [11], [30], [31], [32], [33], [34]. Such requirements are not limited only to traditional CPU or indeed GPU processing elements, but can also employ arrays of reconfigurable FPGA accelerators [35], or even optical computing paradigms [36]. However, they are often mapped onto very different *physical* structures. Taking an inherently 3D (three-dimensional) computing topology and mapping it onto, for example, a fat-tree physical architecture, presents challenges and compromises for cost and scalability alongside their advantages. This indicates that there is merit in building HPC systems as truly 3D physical structures, provided that they can overcome the key design challenges that this presents.

Meanwhile, a substantial body of work has accumulated which addresses the concept of mapping common logical topologies onto uniform and semi-uniform 2D and 3D grid arrays in the physical domain [6], [16], [25], [26], [28], [37], [38], [39], [40], [41], [42], [43]. The ability to 'fold' complex structures such as tori onto both 3D arrays and other physical topologies effectively means that an array structure such as PTCA can accommodate a wide range of computing models and offer some potentially unique advantages alongside its limitations.

All of these cases illustrate similar underlying challenges in terms of IO connectivity, power networks, and cooling pathways. Efficient mapping of 3D logical problems onto true physically 3D-connected hardware architectures, as 3D grids or otherwise, is therefore an important and timely area for investigation if the potential for future systems evolution toward, what are presently considered unconventional paradigms, is to be fully explored.

Another observation can be made of those systems and in general terms: most computing arrays have power grids, IO, and heat extraction systems 'on the outside'. This is observable in the cooling systems of the Google TPU arrays, and the IO cabling of SpiNNaker. It may seem an obscure question, but what if this concept was to be turned 'inside out', so that the power grid, the IO network and the heat extraction are in some senses *on the inside* of the devices which are being attempted to aggregate into arrays instead? The PTCA concept allows us to do exactly that. Indeed, these properties are a necessity for effective 3D tiling and as a result of those requirements some very interesting possibilities arise.

C. POLYHEDRAL TILED COMPUTING ARRAYS

PTCAs are a potentially disruptive and unconventional paradigm for true 3D system assembly, which may address the challenges highlighted in the previous subsection in new ways based upon the principles of the mathematical characteristics of polyhedra (geometric structures capable of tiling in 3 dimensions). Adopting this theoretical foundation and applying those principles in the context of engineering choices required to construct such modules as tileable computing elements, leads us to the concept of the PTCA as examined in this paper.

PTCAs consist of a multiplicity of computing modules which, by virtue of their 3D shapes, can be abutted in one or more dimensions to create arrays. Such 3D geometrically defined shapes are known as polyhedrons. Earlier work investigating practical aspects of this concept focused upon planar hexagonal tiles (HexTiles), which could be assembled into multi-dimensional surfaces, including the truncated-octahedron form [1]. However the most recent work has focused upon complete polyhedral modules with three-dimensional properties, and particularly the truncated-octahedron, which was recognized in work by Lord Kelvin (termed as Kelvin bubbles) as a highly efficient 3D packing shape [44], and is a singular permutahedron which may be uniformly tessellated indefinitely in any plane. The Weaire-Phelan scheme [45] is known to offer a provably better solution to Kelvin's problem, by using multiple different polyhedra. For convenience, truncated-octahedrons, when embodied as computational modules, are forthwith referred to as 'Kelvin Cores' or 'K-cores'. Both Kelvin and Weaire-Phelan forms can be envisaged as 3D tileable computing arrays, but K-cores require only a single core shape, which is advantageous for modular tiling.

Polyhedra are, of course, a widely known concept in mathematics, and there are also notable parallels in other fields on this topic with overlapping theoretical concepts as diverse as 3D tomography, theoretical geometry, and networks [37], [44], [45], [46], [47], [48], but a practical embodiment of such systems in the form of physically composable modular computing arrays has yet to be significantly explored. Part of the motivation for undertaking the work and research reported in this paper is to close that knowledge gap with the use of both theory, modeling, and practical engineering insights.

In hardware terms, a PTCA module will typically contain processing element(s), IO, and data storage components, with options including CPU, GPU, FPGA, TPU, SSD, neuromorphic processor, multi-ported shared memory, or indeed a combination of several elements, or perhaps even emerging technologies such as optical, quantum, or highly unconventional devices. Any of these module types can be combined to achieve heterogeneous arrays of high complexity. Nodes could also be selectively placed to act as local power reservoirs to enhance power distribution with a dynamically varying workload distribution, or as hub nodes within a grid to optimize traffic management, or even as assisted cooling nodes to improve cooling network behavior.

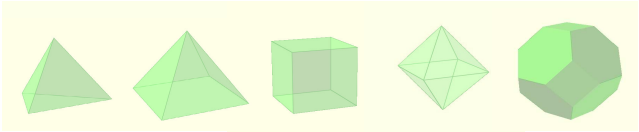


FIGURE 1. Lower order polyhedra. Left to right: tetrahedron, pentahedron, cube, octahedron, truncated-octahedron. Shapes are shown semi-transparent to expose geometry.

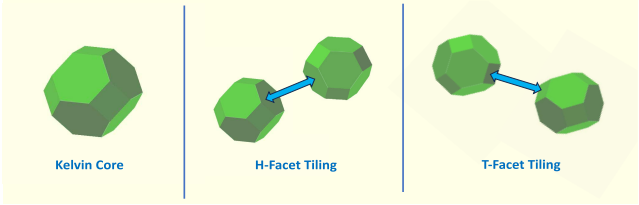


FIGURE 2. 'Kelvin Core' and facet tiling modes. Polyhedral shape (left), H-facet Tiling (middle) and T-facet tiling (right).

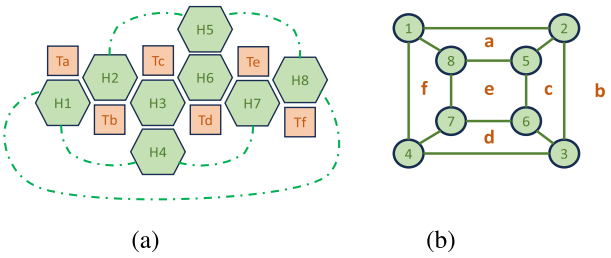
The full spectrum of regular and irregular polyhedra that might pack into some array-like structures is far too large to properly consider here. However there are some obvious choices that might be considered. Regarding our first question (core shapes and sizes), Figure 1 shows arguably the most obvious choices to consider as candidate polyhedra, including square and triangle-based pyramids, cube, diamond and other more complex shapes, and the truncated-octahedron. Figure 2 shows an example of the tiling modes of the truncated-octahedron.

The shapes presented are members of the sets of Platonic and Archimedean solids, with various efficiencies for packing into volumes of space [47]. In part, the answer is that a multitude of shapes can potentially be tiled in some respect [49], but how complete or efficient that packing may be, and how uniform that tiling occurs, is of great importance for genuinely practical systems assemblies. Additionally, the resulting physical topology of a particular packing scheme can be of great importance to the optimal mapping of logical topologies onto its physical structure. Highly incoherent or multivariate packing schemes are likely to be problematic in that respect.

Considering the suitability of these polyhedra as potential PTCA module candidates, and referring to Table 2, it can be observed that only the cube and truncated-octahedron provide rotational symmetry in the class defined here as 'simple', meaning that they can be uniformly packed into an infinitely large 3D array with only a single uniform orientation throughout. Pyramid and diamond have more complex packing attributes, which could/should require rotations of the shapes to achieve tiling and with tessellations at multiple scales. Thus, their base polyhedra form groups of larger and/or more complex polyhedra that are themselves tileable. For instance, some pentahedra may pack together to form rhombic dodecahedra, and these in turn may be tileable 'meta structures'. Although this does not prohibit their validity for PTCA solutions, this complexity is considered

TABLE 2. Basic properties of low order polyhedra.

Polyhedron	F-types	Neighbors	Symmetry	Packing
Tetrahedron (Pyramid)	1	3	complex	0.78
Pentahedron (Pyramid)	2	4 (1+3)	complex	0.95
Cube	1	6	simple	1.00
Octahedron (Diamond)	1	8	complex	0.95
truncated-octahedron	2	14 (6+8)	simple	1.00



(a) (b)

FIGURE 3. Kelvin core facet footprint and edge adjacency graph. This diagram represents (a) the unfolded form of the truncated-octahedron (Kelvin core), where trapezoidal T-facets and hexagonal H-facets have mutual adjacency. When this is translated into (b) a graph representation, it is possible to see how each facet adjoins its neighbors, and thus how each facet might communicate internally with other facets comprising the PTCA core module.

less convenient and potentially creates additional symmetry issues for the layout of facet connectors and makes array assembly a non-trivial task.

It is also apparent that some of the candidates have one facet type, while others have two or more. Notably, the simple cube, with six trapezoidal facets has many of the same properties as the corresponding six trapezoidal facets of the truncated-octahedron. This is of particular convenience as it permits most formulae relating to single packed array T-facet properties to be common to both polyhedra and also permits direct comparisons between the simple cube array forming a standard 3D mesh topology, a truncated-octahedron array forming the same topology in one instance and forming a partially bypassed 3D mesh in another tiling mode, as will be detailed later.

The neighborhood of each polyhedron, the number of other modules with adjacent facets when packed in a grid, indicates that the cube and truncated-octahedron are also more desirable in terms of a grid connectivity scheme with 6 and 14 unique neighbors respectively.

From an engineering point of view, when the facet footprint is 'folded' into a polyhedron, all facets are able to connect to all other facets internally via internal circuit pathways, which could be formed by internal PCB tracks. Figure 3 shows the facet adjacency graph and unfolded facet footprint. Figure 3(a) shows which H-facets are mutual neighbors, and the related adjacency graph in Figure 3(b) shows the connectivity between all hexagonal facets in the truncated-octahedron, which forms a $(2 \times 2 \times 2)$ cube. It should be clear that the internal structure could therefore adopt a common power 'ring' which connects all power delivery facets to a common rail. This ensures that every core

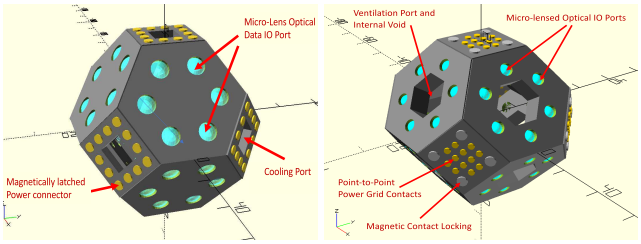


FIGURE 4. Kelvin cores with T-facet and H-facet venting schemes. These renderings show optical IO connections (the blue 'lens' areas on the hexagonal facets), but could utilize pogo pin or other IO data transfer mechanisms. Cooling vents represent entrances into the internal voids within the modules, which facilitate through-flow and internal thermal transfer from hot internal components into the circulating cooling medium (typically being air or liquid based). The number of IO ports and power pins is dependent upon the size of each facet within which they are located.

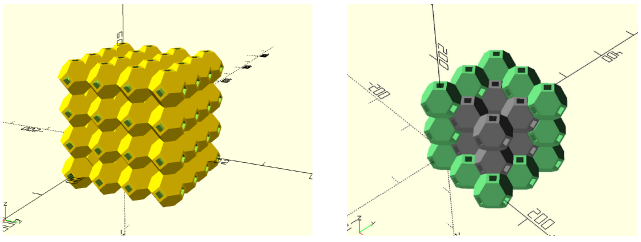


FIGURE 5. Example packing schemes as reported in [4]. (a) single packed grid (left) and (b) double packing example with a cutaway view (right) where gray cores are interleaved with green cores.

can act as a pass-through node to form collective power grids when composed into arrays, and can do so without any external supplementary power wiring or interconnect structures. In effect, the 3D module-to-module power grid is formed entirely by metal structures equivalent to a power bus or ring, but which are entirely internal to the modules. This reflects the earlier comment about PTCA having the advantage that cooling, IO, and power grids are all formed simply by abutting and tiling multiple modules into arrays.

It follows that the truncated-octahedron (K-core) in the envisaged PTCA is a highly relevant candidate for densely packed modular computing arrays. It also has two different facet styles which naturally tile together, as shown in Figure 2. There are 8 hexagonal and 6 trapezoidal facets (referred to here as H-facets and T-facets respectively). The primary focus in the remainder of this paper is the truncated-octahedron, and analysis of the characteristics and capabilities of such a module as a tileable computing element is also considered in terms of technical feasibility, with implications for potential performance.

1) FACET ADJACENCY AND PACKING SCHEMES

Connectivity between nodes is achieved where facets of individual polyhedra coincide and abut with each other in the packed grid. Figure 4 shows several conceptual variations of a Kelvin core, using different facet interface arrangements. The T-facet may provide both power and data connectivity via

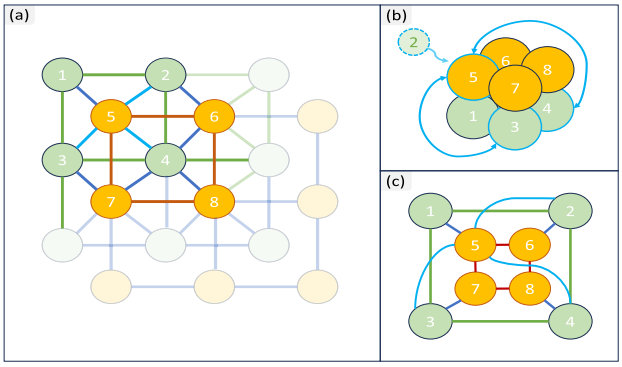


FIGURE 6. Isomorphic equivalences. Case for a double packed truncated-octahedron array horizontal slice, where (a) shows a subsection of one layer of cores (green) and its interleaved double-packed layer (orange), (b) shows the isomorphic equivalent (2 × 2 × 2) cube, and (c) represents the node adjacencies.

abutment of adjacent cores in a point-to-point arrangement or operating in a pass-through bus arrangement, forming a global bus or a localized bus segment. H-facets can also operate in this mode. It is potentially valuable to consider specialization of facets, such that T-facets provide one function and H-facets another. For example T-facets are ideal for power network connection, and some forms of semi-global routing. H-facets offer potential for highly connected data channels. As noted later, a third role of T- and H-facets relates to choices for system cooling.

The tessellation of Kelvin cores is uniform, and results in two possible packing scenarios, as shown in Figure 5: scenario (a) involves only T-facet tiling, arranged in a *primitive cubic* mode, referred to here as a single-packed array, and (b) uses both T- and H-facet tiling in a *body-centered* mode, (also known as a bi-truncated honeycomb) to pack additional Kelvin cores within the spaces created by an initial single-packed array. This is possible because the spaces between cores in the single packed scheme are the exact inverse of the volume of the Kelvin core itself, allowing complete packing. This structure equates to a skewed 3D mesh with additional hyper-connectivities, as visualized in Figure 6.

Notably, this structure (extended into 3D) also incorporates the hypercube topology as an isomorphic sub-graph. This can be extrapolated from the 2D Figure 6(a) where the green and orange nodes (labeled 1 to 4 and 5 to 8 respectively) form the two base squares of a hypercube, and with duplicate sets of nodes extending into layers above (assume corresponding labels 11 to 18) adding an identical pair of squares, linked by an associated vertical edge set such that node 1 connects to node 11, node 2 connects to node 12, and so-on. Thus two hyper-connected cubes are formed. It can be observed then that for the truncated octahedral array, it is possible to form extensible cube or hypercube arrays of cores that are themselves able to operate as 8-hex-facet cubic arrays.

More generally, there are four modes in which this dual-facet polyhedron can operate in a connective tiling:

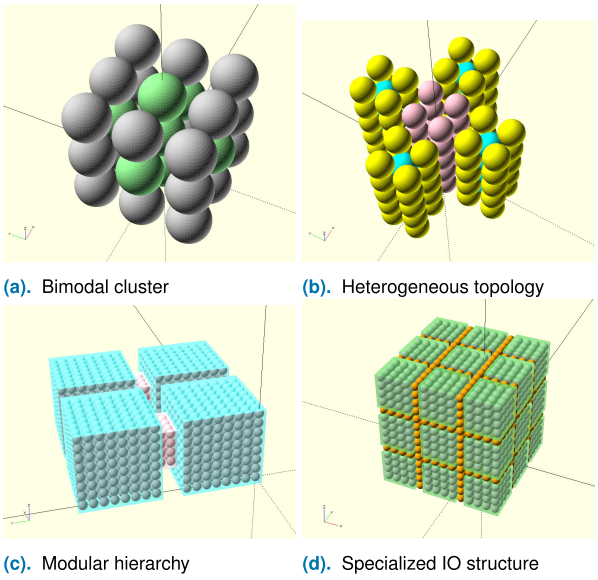


FIGURE 7. Examples of modularity and heterogeneity, where spheres represent kelvin-cores in these visualizations. Core dimensions are a design question, however assuming a quite conservative $d = 100mm$, array dimensions would be as follows: (a) 30 × 30 × 30*cm*, (b) 60 × 60 × 50*cm*, (c) 180 × 180 × 80*cm*, (d) 140 × 140 × 140*cm*.

- A single-packed array using only T-facet interfacing, results in each core having six immediate connected neighbors. This is also directly equivalent to an array of simple cube-shaped modules.
- Double-packed arrays which restrict facet use only to T-facet connectivity would result in two interleaved but independent 3D meshes.
- Double packed arrays with only H-facet interfacing yields eight unique neighbors per core and a single unified grid.
- Double packed arrays employing both T- and H-facet interfaces, with up to 14 immediate unique neighbors per core, and forming a single unified grid.

Arguably, the single-packed Kelvin core arrangement has the same properties as a modular cube, packed to the same grid. However the Kelvin core case has significant space between nodes, and this forms a cooling void in the same way as abutted cores with internal flow vents, providing an additional cooling network (or an alternative if the cores are not themselves internally vented). It is also useful to note that the truncated-octahedron may pack in an intermediate mode where double packing of nodes can be selectively inserted across the grid according to custom computational criteria where machines are being built with highly specific applications. As noted in the next subsection, the internal functionality of each of the interleaved nodes may be one of many heterogeneous possibilities. There may simply be more primary compute nodes, configurable logic based accelerators, intermediary nodes primarily providing enhanced traffic distribution and management, redundant nodes to improve resilience, or indeed other custom purposes.

2) MODULAR HETEROGENEITY

A significant advantage of the PTCA paradigm is the ease of use of heterogeneity in array composition: a design factor emphasized as important by Becker et al. [14]. Tiled modules with externally standardized interfaces can thus have any desired internal specialization, but retain interchangeability, allowing both monolithic or heterogeneous array structures to be rapidly assembled with minimal custom design overheads. Likewise, the physical structure does not have to be as uniform as a cubic arrangement. Consider the illustrations in Figure 7, which show a few examples of heterogeneity, where different core colors represent types of specialized cores such as AI accelerators, GPUs, FPGAs, CPUs, SSDs, memory banks, neuromorphic cores, etc.

Figure 7(a) shows a bi-modal cluster: a combination of two core types in a small array which may represent two interspersed clusters of different device types (e.g. shared memory and CPUs), and could be a sub-grouping repeated across a larger array. Figure 7(b) shows a combination of custom topology and heterogeneous core types, and illustrates another key capability of PTCA systems: whereas cubic arrays are straightforward and in some terms 'general purpose', physical topologies can be constructed that place cores only where they are needed. They are therefore able to reflect aspects of the workload organization for one or more cases within a class of applications. In this case there are many permutations, the pink modules could be an SSD array, the blue modules could hold shared memory columns, and the yellow modules might be GPU nodes. Figure 7(c) highlights the ability of medium to large cubic PTCA arrays to be aggregated into larger systems, with modularity allowing scalable expansion and improved maintainability options. Finally, Figure 7(d) shows an array of sub-modular clusters (colored green) interspersed by planes of nodes envisaged to provide and manage a high-bandwidth many-channel data path between the modules (colored orange), and perhaps employing high radix cross-bar routing ICs at key positions within IO-plane intersections. Effectively this follows a 'many and moderate' rather than 'few and fast' data channels model.

It should be apparent that a complex hierarchical system can be composed in 3D with ease with the PTCA paradigm, a goal that would be more difficult at larger scales in a standard 2D rack-mount approach without a significant design effort around PCB design, bespoke and complex backplane IO arrangements, and other related design trade-offs. PTCA therefore rapidly facilitates a significant diversity of such heterogeneous systems.

3) SYSTEM PROGRAMMABILITY

An obvious question with any kind of array structure might be – *'Is that array's programmability achievable with existing approaches, or do new approaches need to be devised?'*.

It can easily be demonstrated that a single-packed PTCA array is a direct representation of a 3D mesh array, after all,

this is exactly what the structure represents if only the T-facets of neighboring modules abut to each other, and is then identical to an array of cubic modules arranged in the same 3D mesh structure. Since 3D mesh array programmability is widely used and understood as a logical topology, there is no need for new techniques to be developed; they already exist for problems that demand a 3D mesh logical topology solution space.

Folding and mapping techniques are well established for the translation of other logical topologies into mesh-array structures, and much work has been done on mapping structures such as tori into 3D processing arrays with guaranteed scalability [38], [39], [40], [41], [42].

When using double-packed arrays, it is still possible to demonstrate an equivalent 3D array topological structure for programmability. Consider again Figure 6, where it can be seen that the two interleaved array layers that form a double-packed array also create a 3D mesh: a 2D mesh slice is shown in Figure 6, but if this is duplicated in layers above and below, it becomes a 3D mesh. Again, with pre-existing models for programmability of logically 3D mesh topologies, there is no need for new programmability algorithms to be devised to exploit this structure.

Nonetheless, a double-packed array provides additional local connectivities between nodes that are not present in a standard 3D mesh – the hypercube connectivities mentioned earlier for example. This suggests that an enhanced approach to programmability could be beneficial, using existing 3D mesh methodologies but extending this to make optimal use of the links that permit localized packet routes, congestion, and hop-transitions to be bypassed and shortened. That work is beyond the scope of this paper, but quite feasible.

4) FAULT TOLERANCE AND MAINTAINABILITY

With regard to resilience, a key question is how to manage node failures in a large scale array. The first thing to consider is that when scaling to very large array sizes, a strategy that modularizes sub-groups of nodes, such as that illustrated in Figures 7(c), permits the down-time maintenance periods arising due to node failures to be typically limited to only the affected sub-group rather than the entire array, although this is still likely to be an inconvenience in general terms, and thus not a solution in itself.

In a rack-mount system, a node failure typically requires a rack module to be removed and swapped. Though an easy maintenance task, it also means potentially many nodes being removed and in some cases entire multi-node boards being scrapped when only a single node has failed. This can be ameliorated through socket-mounting of critical components but this adds cost and physical volume - there is a tradeoff here between extra up-front cost versus failure-rates and economic maintainability which needs to be balanced.

A strategy that could be exploited for PTCA nodes is that with circuit component positioning on the inside of each H-facet, as visualized in Figures 8 and 9, each node can potentially host multiple processing elements. Thus a degree of redundancy may be incorporated such that extra component cost can be traded off against reduced failure rates. For instance, with eight H-facets, a node might host four primary processing elements and four backups, making total node failure much lower in probability. Alternatively a node might use eight processing elements by default but throttle back workload capacity if individual processors fail. Designating perhaps six active processors as 'normal' operating capacity would maintain apparently normal operation in the presence of slowly accumulating failures.

A typical failure rate for many mainstream processors is often around 1 failure in 250,000 hours. For example, the Intel S1200V3RP Server Board has a MTBF (Mean Time Before Failure) of well over 350,000 hours, and the NVIDIA A100 GPU is anecdotally reported to have a MTBF of between 100,000 to 200,000 hours.

Given a constant failure rate λ for some unit time t, and also that the probability of failure $P_{fail} = (1-e^{-\lambda t})$, it is possible to calculate how many failures will occur in a given period. Suppose that a system of 8,000 nodes is to operate quarterly maintenance schedules, and that each node in the system has a single processor with a MTBF of 200,000 hours. The probability of a node failure, P_{fail}, and the number of incidental failures during each quarter will then be as follows:

$$\begin{aligned} P_{fail} &= 1-e^{-[\frac{1}{200,000}\times\frac{8,760}{4}]} \\ &= 0.01089(\approx 1.1\%) \\ failures\ per\ quarter &= 0.01089 \times 8,000 \approx 87 \end{aligned}$$

This is approaching one failure per day on average, which is far from desirable for any HPC system, even if hot-swap racks are employed and the workload is able to accept and conveniently recover from such failures. However, reliability can be vastly improved by exploiting redundancy techniques.

Consider a PTCA node with a simple 1:1 active redundancy strategy, such that each node contains two processing elements, one of which is normally operational and the other is a normally idle backup device. It then becomes the case that both processing elements in a node must fail in order for that node to become non-operational - a much less frequent event in any given time frame. Assuming the same quarterly maintenance interval, the number of failures can now be determined as follows:-

$$\begin{aligned} P_{fail.dual} &= (P_{fail})^2 = 0.0001185 \\ failures\ per\ quarter &= 0.0001185 \times 8,000 \approx 0.95 \end{aligned}$$

Therefore, the probability of one or more node failures between quarterly maintenance cycles can be substantially reduced, though it can never be completely eliminated. Recent and highly-relevant work by Takanami describes a number of strategies for tolerating failures in 3D grid arrays [43]. Such 'restructuring' and resilience strategies might include the following scenarios in the context of the described PCTA examples given in this paper:

- One, two or even more node failures may be tolerable for extended periods by redistributing workload to other nodes, thus allowing for maintenance to be scheduled at convenient and less disruptive intervals.
- Since every internal node in a Kelvin-core PTCA array has 14 direct neighbors, even a full-node failure event might be patched in a highly localized fashion by using one of those 14 neighbor's spare processors to take on the displaced workload.

It can reasonably be concluded that, although PTCA systems do not have the physical-maintainability convenience of standard rackmount systems based upon schemes such as fat-tree router networks, PTCAs can nonetheless be operated at large scales reliably with suitable design for redundancy.

Clearly, designing appropriate redundancy strategies and modes for continued but degraded performance are essential aspects of any large scale processing array. In the case of PTCA it is particularly important as the effort involved in swapping out faulty nodes is necessarily more complex. The cost-benefit trade-off therefore must balance maintenance and downtime costs against the cost of duplicating components for redundancy. This trade-off may well pivot around systems of a certain size and scale, as well as the level of heat stress and workload regime encountered in use. This is an area of work that could be fruitfully explored and modeled as a 'road map goal'.

The potential to enhance maintainability through hybrid architectural approaches is also noteworthy. A cluster of $(8 \times 8 \times 8)$ node 3D sub-arrays, for instance, might be interconnected to other sub-arrays via more traditional fat-tree routing infrastructure for example, permitting each sub-array to be taken in and out of service relatively easily or even soft-swapped for a backup sub-array that is already in situ. This does of course reduce the full advantages of direct point to point connection in a uniform and contiguous 3D grid, but for certain applications this may be a desirable trade-off for enhanced maintainability.

5) COOLING GRIDS

As noted previously, one question this paper seeks to evaluate is whether power density and cooling is viable in a PTCA. There are two aspects to this:

- *How can cooling be facilitated?*
- *Which range of power densities can be accommodated?*

Cooling issues for tileable modules are fundamentally related to the configuration of a chosen module's facets and achievable packing regimes. Where packing schemes leave spaces between cores (as in the case of singularly packed K-cores), then a continuous cooling flow network is formed by that space. This might be adequate in its own right to provide a volume of air or fluid flow to achieve desired cooling demands. However, for the doubly packed K-cores, and indeed simple cubes, there are no spaces between cores. Instead each core will have an internal spatial void and external 'vents' which provide a path through the interior of the core to any other vent or external array surface. Singularly packed cores could also have internal venting if desired.

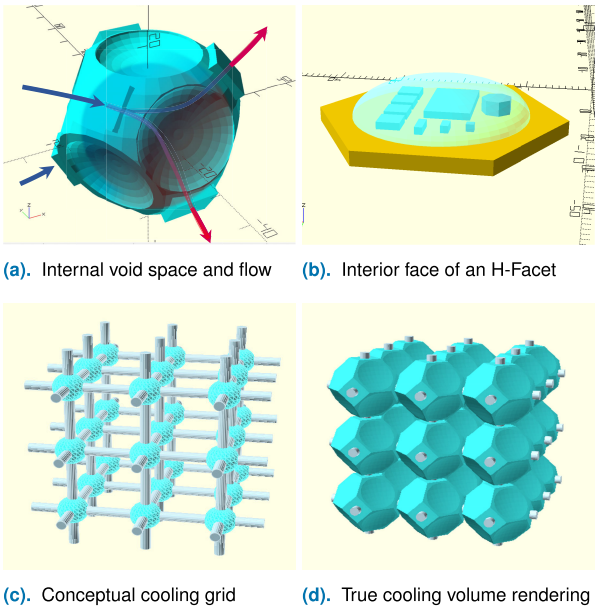


FIGURE 8. T-facet vented core cooling visualizations. Illustrating (a) interior void volume showing example of through-flow. Concave areas represent space occupied by (likely unpackaged) ICs, or MCM modules, located at interior plane of H-facets as shown in (b), where each H-facet can host complex circuitry. (c) Skeleton Cooling grid representation, (d) True volumetric view of cooling void with abutted cores. As for some previous figures, the individual core dimensions are a design parameter, and might feasibly be in the range $60mm \leq d \leq 100mm$.

For K-cores this internal void-space will typically be of the order of 30-50% of the whole core volume for a double-packed scheme, and potentially 130-150% for single-packed schemes with spaces between cores. The vent location is an engineering decision, either T-facet or H-facet locations are suitable, though T-facets are preferred in this paper's analysis, as this means H-facets can host a larger multi-chip-module (MCM) with direct adjacency to external IO interface locations.

An example of the interior void volume formed in a T-facet-vented K-core is given in Figure 8(a), assuming an internal H-facet circuit design concept as shown in Figure 8(b) where components are mounted and bonded in a similar fashion to a Multi-Chip Module (MCM) or System-in-Package (SIP) with a thermally conductive encapsulation and/or heat-transfer structure.

The abutment of tileable cores in an array thus form a continuous internal void flowing through the entire array in multiple directions via T-facet vents, as shown in the topological representation of Figure 8(c). This forms a highly efficient directed cooling network for air or fluid flow with single- or double-packed cores. The true volume rendering is presented in Figure 8(d). Such a volume can exist inside K-cores in both single- and double-packed schemes, and a similar volume can be found between cores in the single-packed case (with or without internal cooling vents).

Such a fluid-flow system would be highly effective for heat removal and useful heat recovery as the PTCA

effectively forms a heat exchange labyrinth. In some senses, this solution sits somewhere between the two extremes of microchannels/microfluidics at the chip level and bulk immersive cooling systems at the other end of the spectrum, with various intermediate systems possible [8], [20], [21], [50], [51], [52], [53], [54]. Advanced implementations might conceivably combine macro and micro-cooling concepts in the same modular structure for example.

6) COOLING PARAMETERS AND MODELING

To help establish the viability of cooling, a critical factor of interest here is the area of the facet flow-vent aperture(s), which have influence over the flow capacity of any externally renewed air or fluid cooling system employed. A vent can be placed at a T-facet or H-facet, and the total area of each of these facets is calculated in the models presented later. This allows a determination of aperture area to be made while accommodating other facet constraints.

Ultimately, scalability in terms of cooling capacity relies upon too many factors to evaluate fully here. Several key considerations are the need for forced cooling flow models, and the impact of surface to volume ratios, as highlighted in earlier work [13]. Interestingly, the variance of surface to volume ratios with scale, once again highlights the relevance of earlier observations about allometric scaling as discussed and cited in Section II-B.

With the models presented, it is possible to determine both the flow capacity of channels, and the surface to volume ratios with a degree of accuracy. Future road-map directions must therefore include the use of fluid dynamics modeling to establish the impact of channel dimensions, flow turbulence, and intersecting flows under varied thermal loading scenarios. Relevant work in the field relating to modeling of complex flow systems and thermal heat transfer mechanisms, at the cabinet, module and component level, will beneficially inform this objective [22], [50], [52], [55], [56], [57], [58], [59], and could be adapted to this novel paradigm.

As a road-map goal, it is clear that a generic cooling model methodology is needed, ideally one which is able to adapt generally to other polyhedral modular compositions, i.e., a generic polyhedral volumetric flow model. This would be valuable as it could integrate with workload modeling to predict power density, thermal dissipation and workload interactions in large grids, not only assisting in identify where existing cooling technologies present limits to PTCA scaling, but perhaps also providing challenges for better future solutions to be conceived. Such models may also inform the possibility to distribute workload to minimize thermal hot-spots within a grid, i.e. thermally aware workload distribution and management.

7) FACET PHYSICAL CONNECTIVITY AND VIABILITY

Another question this paper seeks to address relates to the feasibility of connecting multi-dimensionally tiled objects and the feasibility of suitable connectors. The nature of the facet interfaces is an open design issue with many possibilities. Standard slot-in type connectors would be problematic when composing systems with three axes of tessellation, though not outside the scope of suitable innovation. Being able to slot in a device to other devices on three axes simultaneously with ease is, however, an engineering problem that is as yet not satisfactorily solved.

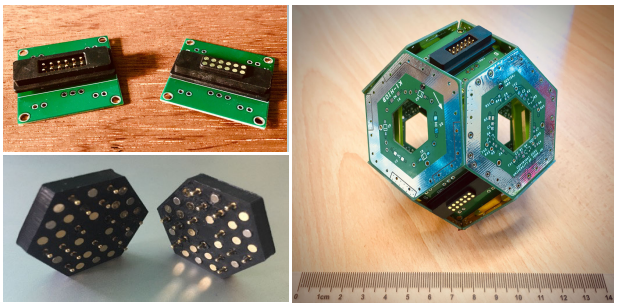
Considering alternatives, another approach is to use multiple parallel pogo-pin connections and magnetic couplings, and this has proven to be highly effective in our experimental work: reducing contact losses due to parallel resistance, multiplying electrical current capacities and reducing voltage losses. The reliability of these type of connectors is high. Manufacturers rate their products in the 10,000s to 100,000s of connect-disconnect (mating) cycles while retaining connection resistance and impedance within rated milliohm tolerances, and this is also corroborated by scientific studies of pogo pin resilience, behaviors under thermal stress, and other factors relevant to a high density system array [60], [61]. An advantage of PTCA is the ability to easily disassemble and reassemble systems, however it is very unlikely that such actions would be repeated many 1000's of times or more. Nonetheless, if necessary, pogo pins can be specified which tolerate high mating-cycle regimes.

It can be concluded that polyhedral modules can connect to others in 3D with practical technology solutions, indeed this approach has been used in a number of prototypes, some examples of which appear in Figure 9, along with a concept for an advanced mass-production assembly methodology.

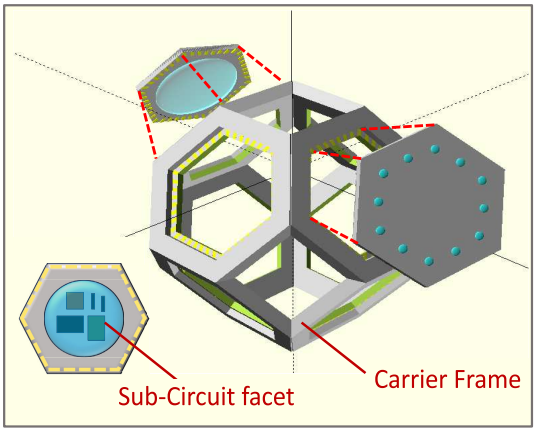
Practical core assembly is an important topic, since it will ultimately have engineering choices that may influence other design aspects. The scenario shown in Figure 9(c) treats H-facets somewhat like PLCC packaged IC modules that can retrofit into a carrier frame. PLCC techniques are well established and could easily be adapted to this form factor, and this would permit any combination of standardized sub-modules to be integrated during manufacture into a standard core frame, allowing a high degree of heterogeneity and specialization even within a single PTCA module.

This approach might use MCMs (multi-chip modules), stacked dies, and other currently evolving techniques. Bare die mounting on MCM substrates and system-in-package (SIP) techniques could allow a high degree of integration and enhanced cooling if encapsulated appropriately and yet would retain the ability to be economically mass-produced in volume.

The carrier frame itself would complete the internal IO pathways and common power grid continuity between T-facets, while the construction of the sub-modules could vary in sophistication from solutions relying upon edge-plated hexagonal PCB's hosting SMD components, through to IC dies bonded to a similar substrate as an MCM assembly. In the most advanced cases, highly integrated VLSI chips housed in hexagonal PLCC chip packages whilst additional optical IO ports and silicon-photonics elements are



(a) pogo-pin connectors (b) K1 Prototype (internal)



(c) Mass-Production Assembly Concept

FIGURE 9. K1 prototype and possible module connectors. (a) Upper left: off-the-shelf connectors mounted on K1-P power boards; Lower left: custom 30-pin H-facet connector plate built by authors (max dia. 28mm: suited to a core of $d \approx 50$mm); (b) K1 prototype modular assembly showing bare module without housing covers $d \approx 90$mm. (c) Production assembly concept.

envisaged as a future evolutions of current chip packaging technologies.

Both power delivery and data IO connectivity may use the pogo-pin arrangement, which in the components shown in Figure 9 occupy only 4mm^2 of facet area each at standard separation. Data IO can also potentially use an opto-coupled arrangement with transmitter-receiver pairs arranged as opposing-facet components, assisted by micro-lens arrangements to aid coupling [62], [63]. A primitive version of the latter concept is used in prototype modules currently being developed as part of a validation project, a photo of which is also shown in Figure 9. Considering the metal-on-metal pogo-pin option, manufacturing data confirms feasibility of pogo-pin data-rates of 10 Gbps or more (e.g., USB 3.2 data lanes), though far inferior to the single-channel 50 Gbps delivered by infiniband HDR, this is pin-to-pin connect at very low hardware cost and can easily be paralleled to many channels if supported by appropriate chipsets.

Meanwhile, recent advances in co-packaged optics, direct-to-die waveguide bonding and integrated on-chip silicon photonics have significant potential while also eliminating issues with multi-lane parasitic signal effects that might be experienced in densely packed wired data lanes. Such technologies are capable of delivering data rates in the many tens of gigabits per second range and with increasing power efficiency, especially for very short range data transfer as envisaged here [63], [64], [65], [66], [67], [68], [69], [70], [71]. A particular emerging technology which achieves very high density multi-fibre IO parallelism, rather than fewer ultra-fast connections, could be very well suited to increasingly small PTCA core modules where high IO bandwidths per cm^2 and low energy per bit are key factors of interest [72]. On the whole, bespoke connector designs are becoming increasingly sophisticated, and there is much more to explore here, where multiple technologies might well be found in a single connector solution supported by a custom interface IC at each facet, combining data buffering/routing, integrated silicon photonics and data messaging protocols in a single 'wrapper' component. Then any desired components can be introduced as part of the core design.

Notably, all of the above techniques lend themselves to mass production at scale, with perhaps the exception of co-packaged optical interconnects, which is nonetheless a rapidly maturing field and expected to become an increasingly economic mass production option in the near future.

When one compares PTCAs to cabinet-based rack and backplane systems, it is clear that PTCAs compose densely packed modules into structures that directly reflect 3D system topologies, while forming data, power and cooling networks simply by the abutment of cores. The viability of such power grids has been validated through experimental prototype hardware test-beds and corresponding simulations of larger scale systems [1]. The feasibility of such systems can be demonstrated on that basis and by the mathematics of the system models to be presented in the next sections of this paper, which will allow a wide range of properties of such systems to be represented.

While PTCA facet IO schemes have an obvious asynchronous point-to-point capability, communications between more distant cores can be facilitated in a variety of ways, as explored in later sections. Facet-to-facet IO schemes may also designate IO pins to support clock domains for synchronous data transfer (e.g. one shared clock pin per abutted facet pair), and/or to support flow control for the whole facet IO channel group. This would allow reporting of the buffer status of the receiving channels to the sending facet and allow hardware flow control with no data loss.

In this relatively unexplored design space, there are clearly many possibilities and equally many unanswered questions. The challenge is that simply building a system is not a practical proposition until we have much clearer ideas about what choices to make in its design, and therefore the technologies and refinements required. Equally, simulations will only tell us how a system will perform based upon how we think we should construct it, or else ignore important questions about those aspects which might be critical for performance.

In this paper, a third approach is added, as described in the next section, using mathematical models to push into the design envelope to a degree that helps to establish feasibility,

TABLE 3. Single packed array characteristics.

Attribute	Symbol	Equation	Eq.
Total Cores	C_{TS}	n^3	(1)
Containment Vol. m^3	V_S	$(n \times d)^3$ [tn 1]	(2)
External Cores [tn 2]	C_{ES}	$n^3 - (n-2)^3$	(3)
Internal Cores [tn 2]	C_{IS}	$(n-2)^3$	(4)
Total T-facets	T_{TS}	$6C_{TS}$	(5)
Total H-facets	H_{TS}	$8C_{TS}$	(6)
External T facets	T_{ES}	$6n^2$	(7)
External H facets	H_{ES}	$24n^2 - 24n + 8$	(8)
Internal T facets	T_{IS}	$T_{TS} - T_{ES}$	(9)
T Bisection factor	B_{TS}	n^2	(10)
Avg Core Power [tn 3]	P_{avg}	$6p \div n$	(11)

[tn 1] d is the distance between two 180^o opposing T-Facets.
[tn 2] for $n \leq 2$, $C_{E^S} = C_{T^S}$, and $C_{I^S} = 0$
[tn 3] p represents power input capacity per individual core T-facet

TABLE 4. Double packed array characteristics.

Attribute	Symbol	Equation	Eq.
Total Cores	C_{TD}	$n^3 + (n-1)^3$	(12)
Containment Vol. m^3	V_D	$(n \times d)^3$	(13)
External Cores [tn 1]	C_{ED}	$n^3 - (n-2)^3$	(14)
Internal Cores [tn 1]	C_{ID}	$(n-1)^3 + (n-2)^3$	(15)
Total T-facets	T_{TD}	$6C_{TD}$	(16)
Total H-facets	H_{TD}	$8C_{TD}$	(17)
External T-facets	T_{ED}	$6n^2 + 6(n-1)^2$	(18)
External H-facets	H_{ED}	$24n^2 - 24n + 8$	(19)
Internal T-facets	T_{ID}	$T_{TD} - T_{ED}$	(20)
Internal H-facets	H_{ID}	$H_{TD} - H_{ED}$	(21)
T Bisection factor	B_{TD}	$n^2 + (n-1)^2$	(22)
H Bisection factor	B_{HD}	$4(n-1)^2$	(23)
Avg Core Power	P_{avg}	$6p \div n$	(24)

[tn 1] for $n \leq 2$, $C_{E^d} = C_{T^d}$, and $C_{I^d} = 0$

TABLE 5. Key core dimensions.

Attribute	Symbol	Equation	Eq.
T-facet side length [tn 1]	T_{side}	$d \div 3$	(25)
T-facet area [tn 2]	T_{area}	$(d \div 3)^2$	(26)
H-facet maximal diameter	H_{dia}	$2d \div 3$	(27)
H-facet side length	H_{side}	$d \div 3$	(28)
H-facet circuit area [tn 3]	H_{area}	$\frac{3}{2} \times \sqrt{3} \times f(d \div 3)^2$	(29)
Whole-core H-circ area	H_A	$8 \times H_{\text{area}}$	(30)

[tn 1] $T_{side} = d$ for modular cube.
[tn 2] $T_{area} = d^2$ for modular cube.
[tn 3] f represents the fraction of the H-facet area usable for circuitry.

TABLE 6. Hop count and neighbourhoods.

Attribute	Symbol	Equation	Eq.
Neighbours (H)	N_H	$(h+1)^3 + h^3$	(31)
Neighbours (T)	N_T	$(2h+1) \times (2h^2 + 2h + 3) \div 3$	(32)
Neighbours (T+H)	N_{TH}	$8h^3 + 6h + 1$	(33)
Max-h (Sngl)	M_S	$3(n-1)$	(34)
Max-h (Dble)	M_D	$m - 1$	(35)

note: *for double-packed cases, m represents the edge dimension for a double packed array, where $m = 2n - 1$, but note that the core count of a double-packed array is substantially larger than single, therefore direct comparisons using n values should be used cautiously.*

practical engineering challenges, and refine the scope for future simulation studies. The expectation is that this allows a path for engineering, modeling, and simulation questions to converge upon a much better understanding of the PTCA concept and its future potential.

III. MATHEMATICAL PROPERTIES OF CUBIC PTCA

In order to assess the properties of PTCA systems, a set of mathematical equations are presented in Tables 3, 4, 5, and 6, relating to cubic arrays of single- or double-packed K-cores.

In a cubic K-core array, the key dimension n represents the number of cores visible across any outer edge of the cube. Therefore a cubic array of size $n = 4$ would have all array cube faces showing 4×4 visible outer cores, just as shown in Figure 5(a). Cuboids with differing values of n in the three key dimensions (height, width, depth) can also be subject to similar analysis, but are not presented here.

Tables 3 and 4 introduce formulae representing a range of system performance properties, but in addition there are dimensional parameters defined in Table 5, and formulae representing connectivity in Table 6. These equations can be used to facilitate first-order determinations of properties such as total system power, power density, data bisection bandwidth, peak connector current, ratios of internal to external IO bandwidth, maximum available power consumption per node, and many more metrics. As mentioned earlier, the single-packed array T-facet properties and equations relate to both the modular cube and the truncated-octahedron K-core, with the exception of T_{side} and T_{area}, where the cube has $T_{\text{side}} = d$ and $T_{\text{area}} = d^2$ respectively.

It might also be noted that containment volumes V_S, V_D represent a cube incorporating the sized array, including wasted space at the bounding edges, since a packing efficiency of 100% is only achievable in an infinite space.

A. FUNDAMENTALS

Some key fundamental properties of PTCA systems include the total number of cores in a given array size, and the number of internal and external cores. These properties are important for some basic considerations such as system size, raw GFLOPS, power delivery capacity and power per node limits.

A more subtle measure is the number of T- and H-facets visible at the array surface. This is the typical external entry point for power and IO connectivity into the grid, and for cooling flows. These equations take into account the location of cores placed at edges, corners and faces of the array.

For example, a single-packed cubic array with $n = 3$ and a relatively large core dimension $d = 0.1$m (10 cm) would have $C_{TS} = n^3 = 27$ cores (Eq. (1)), and occupy a cubic space of $V_S = 30\text{cm} \times 30\text{cm} \times 30\text{cm} = 0.027\text{m}^3$ (Eq. (2)). A double-packed array would also occupy a cube of $V_D = 0.027\text{m}^3$ (Eq. (13)), but has a total of $C_{TD} = 35$ cores (Eq. (12)).

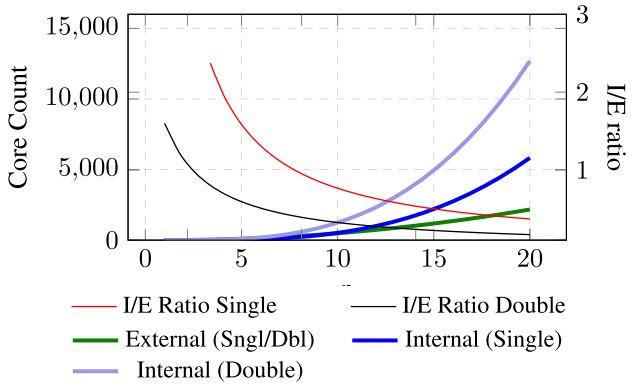


FIGURE 10. Core counts, total, inner and outer cores vs array dimension *n*, for single packing scheme.

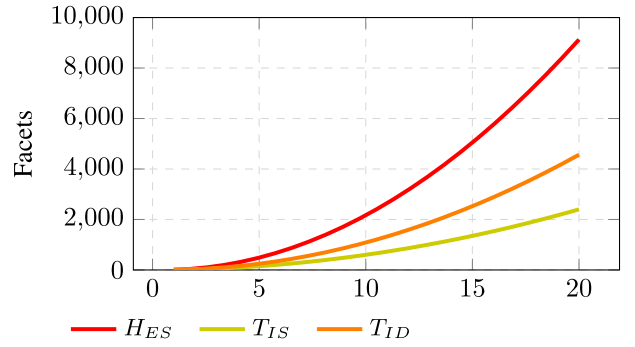


FIGURE 11. External and internal T and H facet counts vs array dimension *n*, for cubic PTCA array.

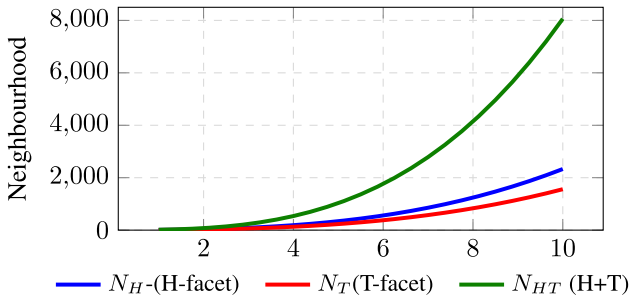


FIGURE 12. Neighborhood size versus number of hops, for T, H and T+H IO tiling.

In this case the core count C_{TD} is only moderately greater than C_{TS}, but tends toward $2 \times C_{TS}$ as n becomes large.

Figure 10 plots the number of internal, external, and total cores, for a range of cubic K-core arrays of external size n, along with the ratio of internal to external cores. These metrics are important for some classes of application, as data bandwidth in and out of the array as a whole is via external core facets, and as the array grows in size the 'surface to volume' ratio decreases.

Equations (26) and (29) identify the inner core surface area available for circuit real-estate. For a core of $d = 80$mm, the H-facet side-length will be $H_{\text{side}} = \frac{1}{3} \times 80 \approx 27$mm (Eq. (28)). If a main circuit area of say 80% of an H-facet is assumed ($f = 0.8$), then the circuit mounting area per H-facet available is determined by Eq. (29) to be $H_{\text{area}} = 1479\text{mm}^2$, or $\approx 14.8\text{cm}^2$. With eight H-facets per K-core, the total circuit mounting area per core approximates to 11×11cm and therefore $H_A \approx 116\text{cm}^2$ (Eq. (30)). Meanwhile the T-facet area would be 711mm^2 (Eq. (26)) compared to an estimated single power pin footprint of less than 4mm^2. If 40% of this area was allocated to a flow-vent, then the vent would have an aperture area of around 17×17mm and flow vent area for the whole core would be $1700\text{mm}^2 \approx 4 \times 4$cm. With an equal portion divided between each of the input and output cooling flows, this is feasibly comparable to cooling pipework diameters of directed cold-plate systems. Adding into this the heat capacity of the flow medium (e.g., air/liquid) and the flow rate, would allow basic evaluations of cooling capacity, as a starting point before moving on to thermal flow dynamics simulations, and to make comparisons to works of related interest [8], [15], [20], [21], [50], [53], [54].

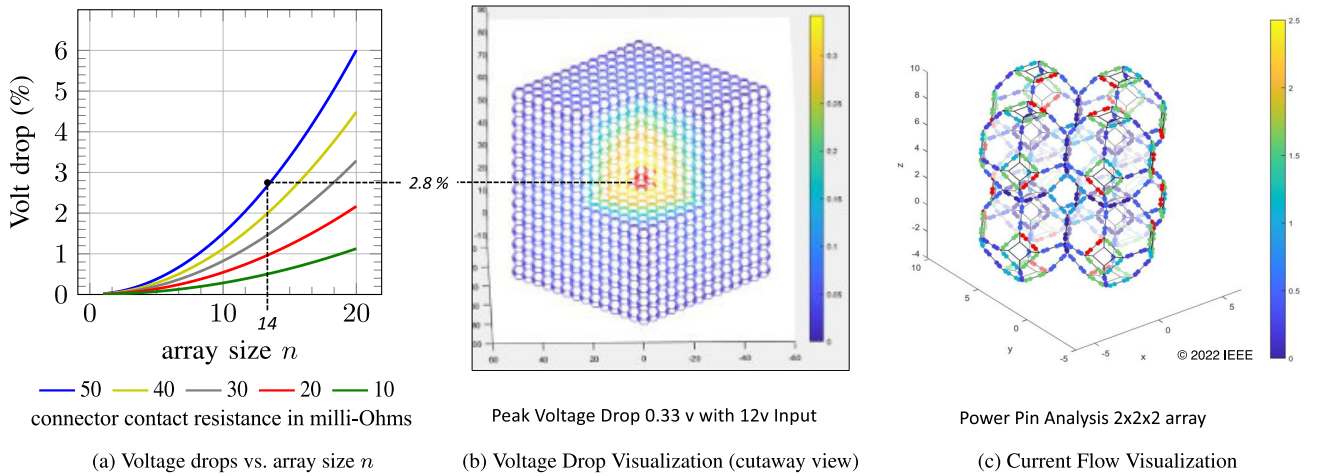
One can see that the basic attributes of a polyhedral module of the types represented here can be assessed and evaluated, and some baseline engineering decisions are able to be informed from this method, including: establishing if enough power pins can be incorporated in chosen facets, estimating adequacy of ventilation flow capacities, and determining overall system size. With such fundamentals established, more sophisticated metrics can now be considered, as outlined in the following subsections.

B. POWER ESTIMATIONS

As stated previously, two questions this paper seeks to answer are whether cascaded power grids can achieve the capacities required of a production-grade system, and whether cooling is viable in such a system. These questions both rely upon having data about core size, power per core, packing mode, and resulting power density that arises. Using the mathematical models introduced, these questions can be answered.

Figure 11 plots internal and external facet counts for single- and double-packing schemes along with the 'surface to volume' ratio. The H-facet count is the same for both single- and double-packing, however the T-facet counts differ, and this may be another reason to favor double-packed systems. Meanwhile, Figure 12 shows the symmetrical neighborhood size for cores reachable from a reference core, within a given hop distance h.

Examining the data it is clear that the neighborhood grows significantly faster as a function of h when the truncated-octahedron provides IO connectivity via both T- and H-facets (14 facets per core in total), as compared to using only H- or T-facet IO (with 8 and 6 neighbors respectively). The T-facet-only case also represents the simple modular cube option as a tileable polyhedron, illustrating the superior connectivity of the chosen K-core model. These equations are important as they are able to be used to determine the external IO bandwidth and external power delivery capacity for a given array size versus the total system core counts. Consider an example where the T-facets



(a) Voltage drops vs. array size n (b) Voltage Drop Visualization (cutaway view) (c) Current Flow Visualization

FIGURE 13. Visualization of projected steady-state voltage drops for $n = 14$. (a) Voltage drops vs. n, plotted for various pogo contact resistances of 10-50 milliohms, (b) Voltage drop visualization for array components in a large cubic array with cutaway view (main image as previously reported in [4], (c) Visualization example of current flow at connector interfaces showing maximum load (red) Image ©2022 IEEE [1].

are used to facilitate power delivery, and the connectors have a capacity of 12 Amps at 12 V (144 W each facet), via a (4+4) pin (3 A, 12 V rated per pin) arrangement (four +VE rail, and four GND). Using Equations (7) or (18) respectively, an array of size $n = 10$ would have the following properties:

- **Single-packed n $\mathcal{D}$ 10**, total of 1000 cores (Eq. (1))
  External T-facets $= T_{ES} = 600$ (Eq. (7))
  Total input power capacity $600 \times 144\text{W} = 86.4\text{kW}$
  Average power per core $P_{avg} \approx 86\text{kW} \div 1000 \approx 86\text{W}$
- **Double-packed n $\mathcal{D}$ 10**, total of 1729 cores (Eq. (12))
  External T-facets $= T_{ED} = 1086$ (Eq. (18))
  Total input power capacity $1086 \times 144\text{W} = 156.3\text{kW}$
  Average power per core $P_{avg} \approx 156\text{kW} \div 1729 \approx 90\text{W}$

Equation (24) provides a first-order figure for average deliverable power per core, P_{avg}, where p represents input power per T-facet. In both cases the average power per core is then able to be determined to be within the range 85-90 Watts.

C. DETAILED POWER GRID MODELING

Note that average power per core, as estimated in previous Section III-B, is a first-order benchmark for estimating core computational capability and limits. A true input wattage delivery per core depends upon each individual core, its power regulator, and its position within the internal power grid, which does of course include interconnection resistances. The effect of these quantities can be determined with precision using a suitable resistance network simulation accounting for pogo-pin resistance effects, the 3-dimensional parallelism of voltage distribution paths, the core power loads, and the core regulator circuit models. This work has already been undertaken and presented in previous work [1], [4], using industry standard SPICE modeling techniques, assisted by automatic model generation tools which derive circuit models from specified array geometry and parameters. Using these techniques, multiple models can be generated to predict behaviors at various scales, given any chosen characteristics of the connectors and other design factors such as power regulator component selection, power load, and so-on.

An example of the voltage drop evaluation using such an analysis is given in Figure 13 where (in that case) the aggregated 3D series-parallel connector resistance network related to connectors, and the on-board regulators of each core are all considered to determine the worst case voltage drop (located at the center of the grid). The simulation case illustrated assumes a modest power connector scheme with two positive and two ground pogo pins per facet with the two groups of pins parallelized to aggregate current capacity per facet of $6A$ and a power capacity of $2 \times 3\,A \times 12\,V = 72W$. An effective parallel contact resistance of 25.0 milliohms was calculated, assuming a worst-case 50 milliohms per individual mated facet-to-facet pair. The worst case voltage drop, with a single packed grid of size $n = 14$ (2,744 cores), is found to be only 0.33 V with a 12 V input (less than 3%), and projected to still only be 6% at $n = 20$ (8,000 cores).

D. POWER DENSITY

It should be noted that a system consuming 156 kW could potentially present challenges for heat dissipation. On this basis, the size of a given system, its total power consumption, and therefore the effective power density of that system, are crucially important. Consider again the example of a cubic array of size $n = 10$, a core dimension $d = 100\text{mm}$, and average power per core of 80 W.

- **Cubic size,** $V_T = [n \times d]^3 = [10 \times 0.1]^3 = 1\text{m}^3$ (Eq. (2))
- **Total cores,** $C_T = n^3 = 1000$ (Eq. (1))
- **Total power** $= 1000 \times 80\text{ W} = 80\text{ kW}$
- **Power density** $= 80,000\text{W}/\text{m}^3 = 8\text{W}/\text{cm}^3$

A not untypical $0.6 \times 0.8 \times 2\text{m}$ 42U rack-mount cabinet has a cubic volume of just under 1m^3. The power density predicted for the case above is thus up to two times a reasonable

upper limit for an air-cooled server cabinet with generic fan cooling (40 kW, or around 4W/cm^3). This particular case therefore indicates quite high power density for traditional air-cooled enclosures, and suggests a requirement to either reduce power per core to perhaps 40 W or the adoption of advanced (but commercially available) cooling solutions such as directed or immersive liquid cooling.

An infrastructural design constraint can also act as the starting point for an evaluation. Consider the 80 W per core case, and assume a cabinet enclosure with a 40 kW upper power limit. This suggests that 40 kW/80 W = 500 cores, and then transposing Eq. (1) yields a single-packed array of size $n = 7$ with 343 cores for a single-cabinet equivalent volume:

$$C_T = n^3, 500 = n^3, n = \lfloor \sqrt[3]{500} \rfloor = \lfloor 7.94 \rfloor \therefore n = 7$$

But, since $7.94 \approx n = 8$, an array of $n = 8$ (with 512 cores) would be possible with a very slight relaxation of power density limits. Of course in practice an array does not have to be uniformly dimensioned. An array of $7 \times 7 \times 10$, with 490 cores, would be just as possible if it suited the application.

A range of system power densities, core wattages, and core dimensions are tabulated in Table 7(a) and 7(b), where Table 7(a) presents data for modular cube and single-packed K-core cases, and Table 7(b) presents data for double-packed K-cores. In terms of the question of power-grid feasibility, it is observable that the distributed grids are achievable with manageable numbers of pins, and low distribution grid voltage drops. Note also that the double-packed case has advantages here at the high end of power density: fewer power grid pins means more freedom for adding T-facet IO pins and/or using smaller cores.

For comparison, in terms of likely core power budgets, Table 8 gives a variety of device power consumption ratings, ranging from a few Watts for an SSD or DDR4 DRAM module to several hundred Watts of a highly specialized and massively integrated device such as the Google TPU v4. Given the trend for power efficiency improvements dictated by Koomey's law, the lower core wattage figures implied for future equivalent computational throughput will be able to support significantly increasing performance levels in the same 1m^3 unit volume and power envelope.

It is important to note that average core power is exactly that: a single packed array at $d = 80$mm and a power density limit of 50kW/m^3 may dictate an average of 28 W to meet the overall power budget, according to data in Table 7. However, with that *average* power rating, some nodes could in fact consume only a few Watts and others 40 W or 50 W. A mix of DRAM banks, SSD modules, and higher wattage compute nodes could be such an example. The quoted 4-pin minimum configuration for that data point, delivers 72 W per T-facet, or a potential total input power of 432 W, ample capacity to permit individual cores to run as high-power nodes while others run cooler. The optimal placement of 'hot' cores to gain the most power efficient power and/or cooling grid placements within an array is also worthy of future investigation, extending some existing preliminary work in this area where evolutionary algorithms have shown promise [4].

TABLE 7. Core counts, power per core, and pin counts for 1m^3 unit volume. Pin-Counts are show as super-scripted figures.

(a) Single Packed Array

d mm	m=n	cores	25kW	50kW	100kW	150kW	200kW
100	10	1000	25^4	50^6	100^{10}	150^{14}	200^{20}
90	11	1331	18^2	37^4	75^8	112^{12}	150^{16}
80	12	1728	14^2	28^4	57^8	86^{10}	115^{14}
70	14	2744	9^2	18^4	36^6	54^8	72^{10}
60	16	4096	6^2	12^2	24^4	36^6	48^8
50	20	8000	3.1^2	6^2	12^4	18^4	25^6
40	25	15625	1.6^2	3.2^2	6^2	9^4	12^4

(b) Double Packed Array

d mm	m=2n-1	cores	25kW	50kW	100kW	150kW	200kW
100	19	1729	14^2	28^4	57^6	86^8	115^{12}
90	21	2331	10^2	21^4	42^6	64^8	85^{10}
80	23	3059	8^2	16^2	32^4	49^6	65^8
70	27	4941	5^2	10^2	20^4	30^4	40^6
60	31	7470	3^2	6^2	13^2	20^4	26^4
50	39	14859	1.6^2	3.3^2	6^2	10^2	13^4
40	49	29449	0.8^2	1.6^2	3^2	5^2	6^2

(c) Core Input/Transfer Power Capacity vs Facet Pin-count

pins	2	4	6	8	10	20
Input Watts	216W	432W	648W	864W	1,080W	2,160W
Pass-thru W	108W	166W	324W	432W	540W	1,080W

Ultimately, therefore the true focus of the question of power density should not be '*is it possible in a 3D tileable array?*', but rather '*what limits arise with particular combinations of modular core size and power consumption per modular core?*' Both of these may be controlled to achieve useful scenarios and their parameters explored by using methods outlined in this paper.

E. DATA IO BANDWIDTH

If, as assumed in the preceding subsection, one were to allocate T-facets as power delivery interfaces, then H-facets might logically be chosen to host data transfer functionality, and this can be evaluated using either Eq. (8) or (9). For example, assume that a single- or double-packed K-core system, at array size $n = 10$, uses an IO interface per facet with a very modest 10 Gbps data transfer rate. The total system external data transfer capacity would be ≈ 22 *Tbps*:

- **External H-facets**, $H_{ED} = 24n^2 - 24n + 8 = 2,168$
- **Bandwidth**, $H_{ED} \times 10$ Gbps ≈ 21.7 Tbps

These figures can be multiplied to more optimistic design cases, for example, if each H-facet had six such IO channels, as in Figure 4, then the external IO bandwidth would be found to be $(6 \times 21.7) \approx 131$ *Tbps*.

This approach can be extended to calculate many of the bandwidth parameters of interest, including total internal data bandwidth and (unidirectional) bisection bandwidth, as illustrated for the case of $n = 10$ in Table 9.

This is a very modest example, using only one IO channel per facet. From recent publications it is clear that

TABLE 8. Approximated device data for selected ICs.

Device and description	Idle	Active	Die Info
M.2 SSD Module	0.3-2.0W	2-9W	
DDR4 256Gb RDIMM	2W	8.5W	
SpiNNaker 18core [19]	0.3W	1.75W	130nm/102mm^2
MDGRAPE4A SOC	-	65W	40nm
Google TPU v4	90W	192W	7nm/600mm^2
Blue Gene/Q A2I 18core	-	55W	45nm/360mm^2
Versal VE1752 AI FPGA	-	50-60W	
Intel i5 12600k 10core	90W	150-220W	7nm
AMD Ryzen7 7700, 8core	20-30W	90-110W	7nm
NVidia A100	-	400W	7nm/826mm^2
96 Core (stacked die) [26]	-	28W	28nm/200mm^2

TABLE 9. Example bandwidth analysis, n = 10, double packed array.

Parameter	Value	Equations/Units
Array size	10	
Total cores	1729	Eq. (12) C_{TD}
Total T facets	10374	Eq. (16) T_{TD}
External T facets	1086	Eq. (18) T_{ED}
Internal T facets	9288	Eq. (20) H_{ID}
T Bisection factor	181	Eq. (22) B_{TD}
Total H facets	13832	Eq. (17) H_{TD}
External H facets	2192	Eq. (19) H_{ED}
Internal H facets	11640	Eq. (21) H_{HD}
H Bisection factor	324	Eq. (23) B_{HD}
External unidirectional BW (T,H)	10.9, 21.9	Tbps
Internal unidirectional BW(T,H)	92.9, 116.4	Tbps
Bisection bandwidths (uni) (T,H)	1.82, 3.24	Tbps

optical IO options offered by the maturing silicon photonics domain are likely to exceed 10-20 Gbps per channel in the next few years at levels of integration, economy, and power use which permit many channels per optical IO module [64], [66], [67], [71]. Likewise, multiple wire-pin connections per facet of the order of 10 Gbps per channel are easily envisaged using Ser-Des converters with relatively low power per channel [73], [74] with some recent HPC systems averaging 30-40 mW per Gbps per channel [69], and equivalent efficiencies approaching 5 mW per Gbps per channel [75]. The pin interfaces themselves are already known to support 10 Gbps data rates (USB 3.1) with differential signaling. Given the very short range of links, range-optimized signaling schemes are certainly also worth future investigation, with bandwidths per channel of the order of 20 or 40 Gbps becoming feasible [76], [77]. Total PTCA external IO bandwidths of 100's of Terabits per second can thus be chosen according to the overall economic envelope of the core modules being designed.

It should be noted that where multiple very high speed wired signal channels operate in close proximity, there are issues such as cross-talk effects between pins. A certain degree of pin-layout optimization can be used to reduce these effects, and they do not prevent feasibility but rather potentially limit the maximum bandwidths achievable per channel. Studies will be required to evaluate these effects in the future and may build on existing work in this domain [78], [79]. It is also valuable to note that, given the typical size of PTCA modules, and their related facets, the pogo-pins related to individual IO channels can actually achieve a significant amount of mutual separation: whereas the minimum might be 2mm between adjacent pins in a densely packed connector, separations of 10mm and more are achievable with moderate populations of IO pins, meaning orders of magnitude reductions in these effects are possible with careful pin placement on facets (for example a 25-fold reduction at 10mm vs 2mm). A model which allows tradeoffs between crosstalk imposed bandwidth limitations versus pin separation and number of pins per facet would allow the optimal choices to be identified in terms of data rate per pin, total number of pins per facet, ideal pin-placement, and thus total optimal facet bandwidth.

F. CONNECTIVITY MODELS

Any practical implementation of a PTCA, with a chosen feasible power grid capacity and power density, must also be able to meet expectations for data connectivity. While bandwidth has been estimated in previous sections, the higher-level question of node-to-node connectivity deserves some attention, particularly as there are capabilities offered by a true 3D tiled array that are not so easily met with other physical array topologies. Some important observations might be made in this direction:

- Single packing requires IO via abutted T-facets, and is increasingly constraining as cores become smaller. It also means that every node has only six direct neighbors and increases worst-case hop-count.
- Double packing is advantageous in that IO takes place via abutting H-facets, giving eight neighbors to any node and permitting diagonal connections within the array. Message paths are shorter.
- Point-to-point communication is the default across the grid. However, the 3D physical locality and adjacency of cores enables useful approaches to reducing worst-case hop counts (as will be shown).
- Physical neighbors permit very short wired datapaths of the order of 10s of mm die-to-die compared to 100s of mm for traditional planar PCB layouts and 2D forms.

The node interconnection schemes in Figure 14 represent the connectivity of cores for single- and double-packed cases, and illustrate the advantages of exploiting a truncated-octahedron rather than a simple modular cube for polyhedral tiling. It can be observed that the shortest path between nodes of worst-case separation (at diagonally opposing corners) is $3(n-1)$ for the single-packed case, since a message must traverse the three X/Y/Z axes in turn to reach the destination. However, in the double-packed cases, where messaging is facilitated via H-facets, the distance is reduced to $2(n-1)$ where n is taken to represent the edge dimension of the primary (outer) grid.

As noted previously, the die-to-die connectivity of neighbor nodes in the PTCA scheme can be achieved at scales

of 10s of mm or less. This means that local IO can be very high bandwidth, and low-energy, compensating for the cost of hops within a large grid (but see also later comments on global channels).

G. 3D ROUTING SCHEME VARIATIONS

Routing cost is always an issue in a 3D grid, being a high diameter topology relative to some others. However, there are numerous opportunities offered by a true 3D packing arrangement to break through some of the limitations of a standard 3D mesh, some of which are outlined below.

1) 3D CUT-THROUGH CHANNELS

Given the short physical distances between near neighbors, and the 3D abutments of neighboring nodes, selected IO channels operating between each adjacent node might be able to be configured as pass-through links traversing intermediate nodes with near-zero decode and routing cost. Rather than data traversing point-to-point with store-and-forward and routing-decision overhead at each hop, a cut-through mode would allow data to pass through multiple hops via a fixed routing, as illustrated in Figure 15, with only a 1 bit addition to latency per stage. Given enough flexibility with this inter-node cut-through capability, such bypass links can even split into tree structures to create persistent rapidly traversed single or multicast data paths between distant node pairs and to create topologically complex low-latency data flows within the grid.

2) LOCAL CLUSTERS

Another possibility is the use of T- or H-facet IO planes to allocate some data channels to support some local channel hierarchies. Addressable routing can be retained, but with the address field being much shorter and thus requiring only to identify the subset of nodes within the cluster. Very low latency broadcast options are also conceivable. Examples are shown in Figure 16, where every member of those clusters will receive a transmitted message within two hops. Overlapping of clusters, or segmentation of an array into adjacent clusters could thus facilitate customized traffic flows, compartmentalization of unrelated traffic, and improved concurrent use of channel bandwidths within grids.

Both cut-through arrangements and local clusters lend themselves to the formation of functionally differentiated locales within larger system arrays. Indeed, a valuable capability of PTCA architectures is the ability to be composed into complex heterogeneous and functionally organized sub-parts without the complexities of designing bespoke PCB racks to achieve physical implementations.

3) X/Y/Z SEMI-GLOBAL CHANNELS

While H-facets support dedicated point-to-point IO, T-facets could provide shared bandwidth semi-global channels with every row or column of nodes in the array providing a single-axis (X, Y, or Z) shared IO channel. Individual or multicast packets could then reach all nodes within a fixed 3-step operation, as illustrated in steps 1, 2, and 3 of Figure 17(a).

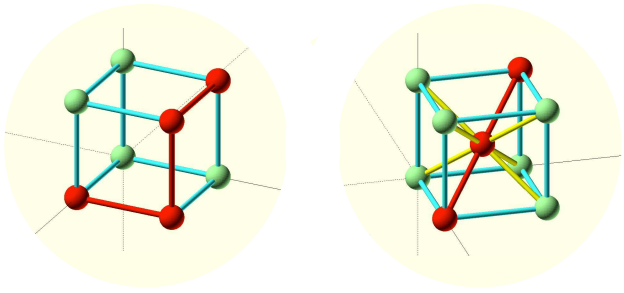
Because the semi-global channels operate on individual X/Y/Z rows and columns they can also facilitate point-to-point routing between any two nodes within 3 'semi-global' steps as shown in Figure 17(b). Given that these channels will typically operate as dedicated routing channels, the global and semi-global routing operations will not impact on congestion or data flow in the general point-to-point network, nor be hindered by that traffic either.

Furthermore these channels can operate independently of others in the same planes, and therefore such operations can occur in parallel with others. The implementation of such a semi-global channel for single, multicast, or broadcast mode could be achieved by a bit-level repeater delay per stage making up a row or column.

4) TORUS WRAPAROUND AND FOLDING

The semi-global X/Y/Z channel scheme can also permit wraparound in all three dimensions without any external/additional wiring. This would be achieved by reserving some or all of the available semi-global channels for this purpose in a persistent fashion. This strategy can support structures of 1D, 2D and 3D tori.

However, suitable folding principles can be applied to map any of those schemes into the single or double-packed



(a) Single Packed Eq. (33) (b) Double Packed Eq.(34)

FIGURE 14. Single and double packed worst-case-pair message routes. Note that each (2 × 2 × 2) case, extensible to ($n \times n \times n$).

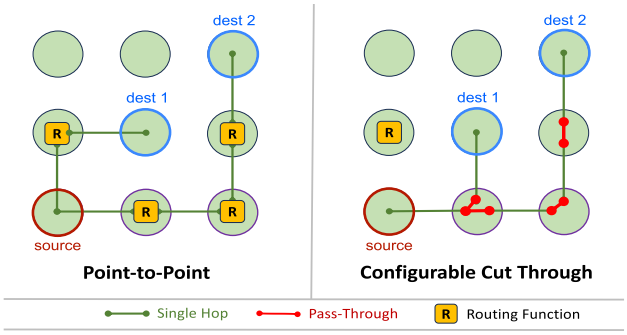


FIGURE 15. Hop count examples, showing (green) point-to-point connections and (red) single level pass-through. Only two dimensions are shown but the same principle applies in 3D. In this example, point-to-point routing requires up to 4 hops with 3 packet routing decision points and related latencies, whereas the cut-through mode requires 4 single-bit latency hops and zero routing cost.

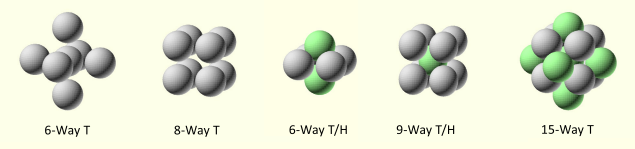


FIGURE 16. Some examples of near-neighbor cluster formations, where the worst case hop-counts are typically 2 hops (3 for 8-Way T).

grid with few or no long links, as explored in various studies [38], [39], [40], [41], [42]. Of these, [41] is a notable study, focusing upon folding solutions in a BlueGene/L HPC platform, which is effectively a physical 3D mesh topology for which the PTCA cases evaluated here could be substituted with the same principles. These techniques can thus ensure that many highly desirable computing topologies can be physically mapped into PTCA systems such that only internal paths formed by the tiling of the cores themselves are involved. Consequently, concerns about high hop-count derived topological edge-to-edge latencies are very capable of being reduced to single hop latency cost in many folded topological cases.

An additional observation is that the double-packed PTCA array actually contains an embedding of a 3D mesh within its interconnect scheme, and therefore is itself a 3D mesh with all of the aforementioned capabilities that apply to the single packed 3D PTCA array. However, since there are additional interconnects available in the double-packed array when viewed as such a 3D mesh, it has bandwidth, neighborhood, and latency advantages over the single-packed case and this presents a form of hypermesh.

The extra connectivities of the PTCA hypermesh may reveal superior folding opportunities as compared to a standard 3D mesh. A valuable contribution or 'road-map goal' in this field would therefore be to explore and review folding methodologies in the context of the double-packed K-core array and contrast these with those applied in other topologies such as a plain 3D mesh. Ultimately a generalized approach that can examine folding of frequently used logical computing topologies with any polyhedral module shape (i.e. not only K-cores) would be a substantial achievement, allowing polyhedral choices to be sifted into those capable of efficiently supporting particular topologies or not.

H. ROUTING COSTS

It is useful at this point to quantify some routing cost factors that could apply to those scenarios and illustrate how they scale with grid size. Some assumptions are needed in order to make an evaluation at this point, particularly the fundamentals of a low-level data protocol for routing. Several modes of operation are outlined in Figure 18.

This is a very primitive data link layer model: assuming a more complex framework at this point, in the absence of a comprehensive system design specification, would be entirely arbitrary, and the intention here is to explore feasibility rather than demonstrate optimality. This model

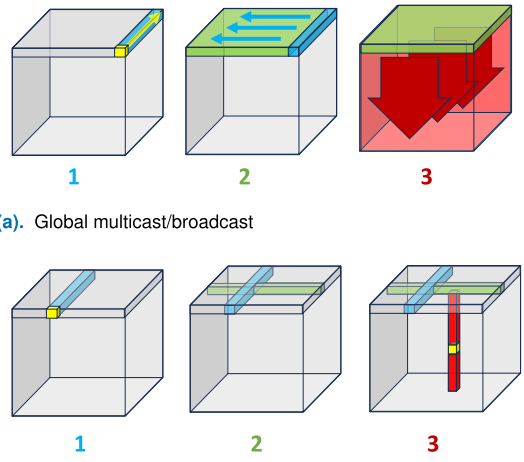


FIGURE 17. Examples of semi-global channel use.

does, however, support point-to-point transfer, semi-global and global routing, broadcast, and cut-through routing and circuit switching (persistent link) concepts.

Destination routing is assumed here to operate on an asynchronous level, with a stop+start bit preamble, immediately followed by a single mode bit dictating local data transfer mode (0) indicating no onward routing required or onward routing mode (1) with an absolute address field, followed by payload data which will be subsequently routed and forwarded by the present receiving node.

Source routing is also represented, where the entire switching decision path is mapped out as a number of successive 4-bit switching codes representing all 14 possible facet choices at the next hop, plus some control states such as 1111_2 representing the end point of an address chain. The switching codes are dropped from the message as they are used, so the message becomes shorter as it progresses. Although this means there is a long series of bits for distant hop destinations, there is no routing table lookup delay at each intermediate stage, just a simple local-port selection.

Cut-through mode represents the case where a point-to-point link is persistently configured as a dedicated circuit-switched link or intermediate stage of a routing path, and only requires the stop-start preamble with no routing address, followed by data. The final stage of such a link can also potentially receive standard addressed packets and act as an onward routing resource with lowered latency.

Semi global mode operates in this case such that nodes in a row or column can only transmit after receiving a circulating token, which that node then transforms into the message header followed by data.

While these modes represent the low-level basic framework, more complex messaging protocols would of course be built upon these foundations according to needs. Using these assumptions it is possible to evaluate raw message latency under various modes of operation, where latency is defined here as the delay between beginning to send the

TABLE 10. Example latencies for array dimensions, n = 4 to n = 40. Assumptions: PTCA IO data rate of 10 Gbps per channel and zero length message payload. PTCA routing decision delay costs are R = 5ns per hop (Destination routing with static routing table) and R = 1ns per hop (Source routing). Fat-Tree case assumes 24-port Infiniband Switches with 130ns port-to-port delay, whilst disregarding NIC HCA RX and TX Latencies for both PTCA and Fat-Tree Infiniband models.

	Unit Delay (1 Unit = 100ps)							Equivalent Delay (ns)						
Grid Size	n = 4	n = 8	n = 12	n = 16	n = 20	n = 30	n = 40	n = 4	n = 8	n = 12	n = 16	n = 20	n=30	n = 40
Total Node Count C_{TD}	91	855	3,059	7,471	14.8k	51.3k	123.3k	91	855	3,059	7,471	14.8k	51.3k	123.3k
General														
Single Cut Through	4	4	4	4	4	4	4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
Worst-case Cut-Through	12	24	36	48	60	90	120	1.2	2.4	3.6	4.8	6.0	9.0	12.0
Global 3-Step X-Y-Z	42	72	101	126	153	216	279	4.2	7.2	10.1	12.6	15.3	21.6	27.9
Destination Routing	A = 7	A = 10	A = 12	A = 13	A = 14	A = 16	A = 17	A = 7	A = 10	A = 12	A = 13	A = 14	A = 16	A = 17
Single Neighbor Hop	61	64	66	67	68	70	71	6.1	6.4	6.6	6.7	6.8	7.0	7.1
Worst-case hop (H only)	366	896	1452	2010	2584	4060	5538	36.6	90	145	201	259	406	554
Worst-case hop (T only)	549	1344	2178	3015	3876	6090	8307	55	134	218	302	388	609	831
Source Routing	A = 4	A = 4	A = 4	A = 4	A = 4	A = 4	A = 4	A = 4	A = 4	A = 4	A = 4	A = 4	A = 4	A = 4
Single Neighbor Hop	18	18	18	18	18	18	18	1.8	1.8	1.8	1.8	1.8	1.8	1.8
Worst-case hop (H only)	108	252	396	540	684	1044	1404	10.8	25.2	39.6	54.0	68.4	104.4	140.4
Worst-case hop (T only)	162	378	594	810	1026	1566	2106	16.2	37.8	59.4	81.0	102.6	156.6	210.6
Fat Tree Infiniband														
Single hop (min latency)	1300	1300	1300	1300	1300	1300	1300	130	130	130	130	130	130	130
Worst-case port to port	3900	6500	6500	6500	9100	9100	9100	390	650	650	650	910	910	910
Levels | Routers	2	3	3	3	4	4	4	6	40	140	338	672	2,322	5,569

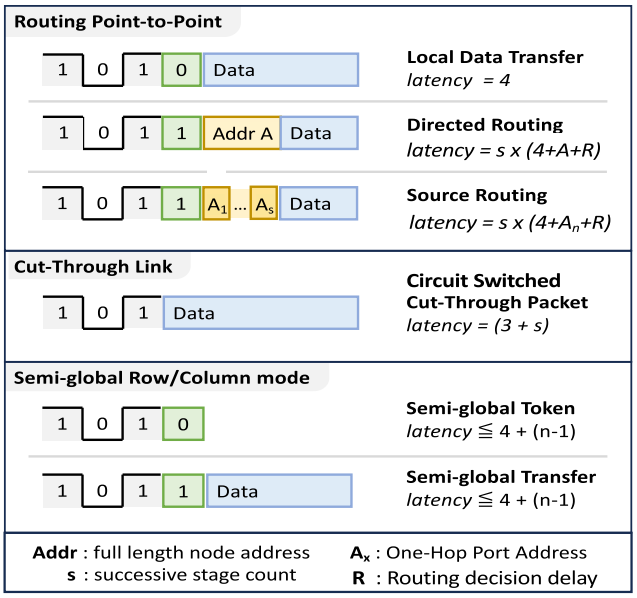


FIGURE 18. A primitive message framework. Message latency equations assume a non-store-and-forward model, with hardware flow control and zero message length (i.e. excludes payload). Directed routing requires a full final destination address bit field *A*, whilst source routing assumes an address field of successive 4-bit onward port selection sub-addresses $A_1 - A_x$. Cut-through links assume a one-bit forwarding delay once a link is established.

message preamble and header, and the destination being ready to receive the first data bit of the payload. For this model, a SerDes latency of 4UI (Unit Interval) is adopted, as described by Han et al. [74] such that a message header length of m bits must round up to a multiple of 4 bits in practice ($\lceil m \div 4 \rceil \times 4$).

Latency estimates in Table 10 show various T-facet and H-facet routing cost factors for a double-packed PTCA grid using the simple routing models introduced, and assumes a zero-length message case with hardware flow control, no store-and-forward protocol, and no congestion. Destination routing and source routing are shown for single hop, and worst-case multi-hop paths via both T and H facets respectively. It is apparent that H-facet routing offers a significant gain, confirming that K-core arrays perform significantly better than arrays of simple cubes (where only T-facets are available). Worst-case routing costs correspond to the most distant node pairs in the grid. Hop-based routing decisions are assumed to require 5ns (50 Unit delay intervals) for destination routing, and 10 unit delays (1ns) for source routing, with the assumption that routing lookup is via a static routing table. Global messaging and cut-through mode have no lookup cost. It is clear that cut-through mode offers a substantial benefit for distant connections, and might be exploited to establish a network of fast data highways to 'short-cut' latencies across large grids, as imagined in the earlier example of Figure 7(d) for instance.

Meanwhile, a comparator case for a simple Fat-Tree router scenario is provided, using 24-port Infiniband™ switches in 1, 2, 3 or more switching levels, according to the total double-packed node count C_{TD}, and assuming switching latency of 130ns port-to-port per switch level. Infiniband switching latencies are frequently of the order of 100-200ns [80], and figures of 130ns port-to-port latency are obtainable in manufacturer data (e.g. Mellanox QM8700/QM8790). In this analysis, NIC HCA TX/RX latencies are ignored, since the PTCA model also ignores any equivalent overhead. The number of levels and the total router count is also estimated assuming a Clos non-blocking

configuration. It is also noted that higher layer routers (from edge to aggregate to core) will require successively higher link bandwidths, which significantly alters the cost of that subset of routers as compared to routers at the lowest levels in the hierarchy. Evidently, the large number of routers required for moderate to large array sizes would be a significant cost factor, and one which PTCA can eliminate. It is notable that router cost is not just an economic measure, but also includes physical space, substantial further overhead for cabling, power consumption and heat emissions. These factors are all relevant to future cost-performance evaluations.

A figure of 10 Gbps is a moderate choice, compatible with a wide range of IO connection schemes. However more specialized (and more costly) IO options can achieve considerably more. Therefore much higher data rates are possible, provided that the digital circuitry and processing elements connecting to them can usefully manage such quantities of data. Whether it is better to have fewer pins at very high data rates or multiple pins of lower bandwidth but equivalent in combination, really depends upon the physical size of the core, the kind of processing elements within them, latency versus throughput balances, and the intended range of application workloads. Future research into higher speed optical links [63], [64], [65], [66], [67], [68], [71], wireless links [3], [74], and very low-power ultra-short data paths would be valuable.

IV. CASE-STUDY EVALUATIONS

Using the equations and analysis presented previously, we can evaluate the properties of a range of different scenarios. These are not intended to be like-for-like topology comparisons or reproductions of the same specific capability, but rather as guides to the capabilities of PTCA systems at similar scales. The cases are:

(a) A 'million core' use-case
(b) A moderate scale with medium-power cores
(c) A one cubic metre system module at $50kW/m^3$
(d) A near-future PTCA Exascale system

Before specific details of these test cases are discussed, some general assumptions are made about data IO, and power:

- Power is delivered via T-facets with the pin-count as defined in each use-case with 3 A, 12 V, per pin.
- Point-to-point IO channels operate at 10 Gbps when using wired pin contacts.
- Optical IO channels are assumed to operate at 20 Gbps.
- GFLOPS/Watt baselines and related performance projections, are based on whole-system power, and power per core is total system power divided by core count.

For each scenario, Table 11 lists the design parameters, while Table 12 provides the resulting system projections.

A. A MILLION-CORE USE-CASE

This scenario is broadly aligned to the SpiNNaker-106 ''million core'' architecture [6], [16] as a reference point for scale and complexity. The SpiNNaker system has a huge number of processing devices but very low power per device, consisting of 10 server cabinets, up to 1,200 rack mount 'SpiNN-5' custom PCB modules each containing 48 specialized 18-core processing units (16 active plus 2 spare cores) optimized for neuromorphic computations. This corresponds to a total of 57,600 multicore units and 1,036,800 total processing elements, while consuming an estimated 90 kW at peak load and thus averaging around 1.7 W per multicore SOC (System on Chip) including IO and peripheral logic power usage (a very low individual node consumption by current standards).

SpiNNaker uses a 240 × 240 hexagonal torus logical topology with node-to-node 250 Mbps links, implying a bisection bandwidth of 120 Gbps. Rack modules interconnect via a custom high-speed multiplexed 'SpiNNlink' data highway at 480 Mbps per link. The custom 48-chip boards are around $53,000$ mm^2, with an area of 33×33 *mm* per chip.

The design choices for this use-case are summarized in Table 11(a), with a rationale as follows: The very low power usage of the 18-core SpiNNaker IC allows an assumption of each PTCA core being able to host multiple multicore devices, in this case assuming eight devices per core with a still modest total power per core of less than 14 W. This then is a question of granularity: how big should a core be and how much hardware should be embedded in each core? A core size of $d = 80$ mm would permit an H-facet circuit area of nearly 30 × 30 mm, which could accommodate an MCM bare-die assembly including a device of similar dimension to a SpiNNaker multicore plus additional components, and is very similar to the $\approx 33 \times 33$ mm average PCB area occupied by the IC-packaged SpiNNaker IC nodes. In terms of scale, the closest equivalent multicore count would then be $n = 20$ for a PTCA system of 4,096 single-packed K-cores hosting 65,536 multicore devices, or an $n = 16$ PTCA with 7,471 double-packed K-cores hosting 62,800 SpiNNaker multicore devices. In the latter case this equates to a total of 1,130,400 processing elements, of which 1,004,800 would be active if adopting the same 16 of 18 redundancy scheme.

The mathematical models suggest that all three cases achieve power densities within the limits of air cooling, with the double-packed PTCA system near the upper limit, at around 47 kWh/m^3. Meanwhile, bisection bandwidths of the order of over 70 Tbps are predicted, with cores having an individual IO bandwidth of up to 0.72 Tbps.

B. A MEDIUM-SCALE, MEDIUM-POWER SCENARIO

For this system scenario, we use MDGRAPE-4A as a source of inspiration, with a modest number of cores and a medium power demand per core. MDGRAPE-4A has a 3D inspired logical topology, typically implemented as an 8 × 8 × 8 logical 3D array of 512 modules consuming around 125 W each (65 kW total system power load), and having 12 × 6 Gbps wired and optical links for each chip-to-chip neighbor [9], [10]. Table 11(b) shows design choices for a system of similar scale and power budget as this reference case.

TABLE 11. Use case design parameters.

(a) 'Million Core' Use Case

Parameter	Details
Core Size	[1] Cube/Single-KC 70 mm [2] Double KC 80 mm
PEs per core	8 (one 16-multicore per H-facet)
Core Power	≈ 14 W$(8 \times 1.7$ W$)$
IO T/H/G (Cube)	$\{4, -, 4\} \times 10$ Gbps
IO T/H/G (Single KC)	$\{4, -, 4\} \times 10$ Gbps
IO T/H/G (Double KC)	$\{2, 6, 2\} \times 10$ Gbps

(b) Moderate Scale Medium Power Case

Parameter	Details
Core Size	100 mm
PEs per core	One equivalent SoC per core
Core Power	$\approx 48/24$W
IO T/H/G (Cube)	$\{8, -, 8\} \times 10$ Gbps
IO T/H/G (Single KC)	$\{4, -, 4\} \times 10$ Gbps
IO T/H/G (Double KC)	$\{4, 12, 4\} \times 10$ Gbps

(c) Single Cubic Metre Array

Parameter	Details
Core Size	80 mm
PEs per core	One equivalent SoC per core
Core Power	≈ 25 W$/14$ W
IO T/H/G (Cube)	$\{8, -, 8\} \times 10$ Gbps
IO T/H/G (Single KC)	$\{4, -, 4\} \times 10$ Gbps
IO T/H/G (Double KC)	$\{4, 6, 4\} \times 10$ Gbps

(d) Future Exaflop System, Liquid Cooling

Parameter	Details
Core Size	80 mm
PEs per core	One equivalent SoC per core
Core Power	≈ 42 W$/22$ W
IO T/H/G (Cube)	$\{8, -, 8\} \times 20$ Gbps
IO T/H/G (Single KC)	$\{8, -, 8\} \times 20$ Gbps
IO T/H/G (Double KC)	$\{8, 8, 8\} \times 20$ Gbps

A single-packed PTCA of size $n = 8$ has the identical number of 512 cores, but results in a high power density that would require liquid cooling. An array size of $n = 11$ with 48W per core is more manageable. A core size of 100 mm is assumed, alongside a comparable core-to-core data rate of 12×10 Gbps per H-facet, or 4×10 Gbps per H-facet for single-packed cases, and a global channel data rate of 2.5 Gbps $\times$ 4 channels per X/Y/Z row/column. Considering the analysis of this scenario, as detailed in Table 12(b), indicates that a feasible solution is possible.

Increasing the size of the K-cores would help to reduce power density, but is not an ideal solution as they are already relatively large in this scenario, and is thus sub-optimal. Another option would be to use a heterogeneous mix of cores, some being of lower power demand. This would reduce the overall power density as part of a design strategy.

However, the processing elements assumed are based upon a mature VLSI process node at 40 nm. Technology scaling can be expected to deliver power efficiency improvements. For example, scaling through technology nodes from 40 nm down to a 10 nm process potentially offers the order of a 5-fold power efficiency gain, and scaling is already well beyond that point in current leading-edge fabrication processes. This indicates that a newer SOC processing element could deliver much better performance within a similar air-cooled PTCA design envelope, and potentially support more SOCs (or more complex ICs) per core. Indeed if this solution was using 10 nm equivalent devices, then it could double the performance of the above use-case while still reducing power density to within manageable air-cooled system parameters.

Extrapolating to lower nm process nodes, toward the current state-of-the-art node at 2nm for example, can be done with lower confidence, but may imply of the order of 8 to 12 fold improvements. The shifting tradeoffs between static and dynamic power across that range makes a robust comparison at this point difficult without more detailed workload models.

C. 50 KWH, ONE CUBIC METRE SYSTEM

The third scenario assumes that the goal is to maximize the computational density of a system in a 1-metre cubic volume within the limits of air cooling. This could then represent a standalone system or a module in a system to be coupled to others as illustrated for example in Figure 7(c). The objective here is to achieve a relatively large number of cores at modest power per core. A suitable compromise can be found when we assume a core size of $d = 80mm$, as detailed in Table 11(c).

This system delivers a raw performance of around 2 PetaFLOPS per cubic metre at an assumed power efficiency of 50 GFLOPS/Watt, with between approximately 1,700 and 3,000 cores, and total bisection bandwidth of nearly 64 Tbps.

D. POSSIBLE PTCA EXASCALE SCENARIO

In this scenario, the analysis assumes a 2030 epoch, where power efficiency is predicted to be in excess of 400 GigaFLOPS/watt (further justification for arriving at this figure is given in Section IV-E). The goal is therefore to achieve a one ExaFLOP system configuration using the PTCA paradigm and that efficiency point. Given the increased computational throughput this would imply, the IO capabilities of this test case are more substantial, assuming a number of 20 Gbps IO channels per facet, with the assumption that optical connectivity is employed.

An important design choice in this system is that the model given in Figure 7(c) is adopted, such that while the dimension of the array overall is $n = 41$, this is populated by twenty-seven smaller cubic arrays of $n = 13$, interspersed by nodes dedicated to forming an intersecting IO plane. These additional nodes are not counted in the contribution to the performance of $\approx$ 1 ExaFLOPS achieved in these design cases, however the same power consumption is assumed for all nodes including the IO plane cases. The chosen configuration is given in Table 11(d).

Adopting optical IO channels results in the Exascale system having array bisection bandwidths approaching 1.3 Pbps in the double-packed case, with single core IO

TABLE 12. Selected TCA evaluation cases.

(a) 'Million core' case	RackMount Case	Basic Tiled Cube	KC Single Packed	KC Double Packed	Eqns/Notes
p_1 Assembly mode	10x42U Racks x 5x4U 4U: 24 pcb x 48 PE	n=16 single packed face-centred-cubic	n=16 single packed face-centred-cubic	n=13 double packed bifurcated honeycomb	
p_2 Core Count , Power	57,600, $\approx$ 1.7W	4,096, 22W	4,096, 22W	3,925, 22W	(1),(12)
p_3 Required Power	90 kW	90 kW (2 pins)	90 kW (2 pins)	86 kW (2 pins)	pins per T-facet
p_4 Available Power	n/a	111 kW (123%)	111 kW (123%)	135 kW (157%)	(7),(18)
p_5 Core Size , System Size	$\approx 6x1x2m^1$	70mm , 1.1m cube	70mm , 1.1m cube	80mm , 1.0m cube	(2)
p_6 System Volume	$\approx 12m^3$	$1.4m^3$	$1.4m^3$	$1.1m^3$	$(n \times d)^3$
p_7 Power Density	$\approx 7.5kW/m^3$	$64.1kW/m^3$	$64.1kW/m^3$	$76.7kW/m^3$	$p_3 \div p_6$
p_8 Core Bw, Tot Bisect. BW	1.5 Gbps , 240 Gbps	480 Gbps , 10.2 Tbps	480 Gbps , 10.2 Tbps	720 Gbps , 72.5 Tbps	(10),(23)
p_9 T/H/G Bisection BW	n/a	(5.1, - , 5.1) Tbps	(5.1, - , 5.1) Tbps	(1.7, 69.1, 1.7) Tbps	
p_{10} Raw PetaFlops @50 GFLOPS/W	4.5 Pflops	4.5 PFlops	4.5 PFlops	4.3 PFlops	(11),(24)
p_{11} H-facets circ area	n/a	n/a	$9,060mm^2$	$11,833mm^2$	(28)
p_{12} Void Vol, per core, total	n/a	187 cm^3, 0.8 m^3	187 cm^3, 0.8 m^3	64 cm^3, 0.3 m^3	at 30% void ratio
p_{13} T-Vent inflow, core, total	n/a	218 mm^2,0.17 m^2	218 mm^2,0.17 m^2	284 mm^2,0.27 m^2	at 40%, (7),(18),(26)

(b) Moderate Scale case	RackMount Case	Basic Tiled Cube	KC Single Packed	KC Double Packed	Eqns/Notes
p_1 Assembly mode	64 x Rack	n=11 single packed face-centred-cubic	n=11 single packed face-centred-cubic	n=11 double packed bifurcated honeycomb	
p_2 Core Count , Power	512 , 125W	1,331, 48W	1,331, 48W	2,331, 28W	(1),(12)
p_3 Required Power	65 kW	64 kW (3 pins)	64 kW (3 pins)	65 kW (2 pins)	pins per T-facet
p_4 Available Power	n/a	78 kW (123%)	78 kW (123%)	95 kW (146%)	(7),(18)
p_5 Core Size , System Size	$4 \times 1.2m^2$	100mm , 1.1m cube	100mm , 1.1m cube	100mm , 1.1m cube	(2)
p_6 System Volume	$\approx 4.8m^3$	$1.3m^3$	$1.3m^3$	$1.3m^3$	$(n \times d)^3$
p_7 Power Density	$\approx 13.5kW/m^3$	$48.0kW/m^3$	$48.0kW/m^3$	$49.0kW/m^3$	$p_3 \div p_6$
p_8 Core Bw, Tot Bisect. BW	"- , -"	960 Gbps , 9.7 Tbps	480 Gbps , 4.8 Tbps	1440 Gbps , 100.8 Tbps	(10),(23)
p_9 T/H/G Bisection BW	n/a	(4.8, - , 4.8) Tbps	(2.4, - , 2.4) Tbps	(2.4, 96., 2.4) Tbps	
p_{10} Raw PetaFlops @50 GFLOPS/W	3.25 Pflops	3.2 PFlops	3.2 PFlops	3.3 PFlops	(11),(24)
p_{11} H-facets circ area	n/a	n/a	$18,489mm^2$	$18,489mm^2$	(28)
p_{12} Void Vol, per core, total	n/a	545 cm^3, 0.7 m^3	545 cm^3, 0.7 m^3	126 cm^3, 0.3 m^3	at 30% void ratio
p_{13} T-Vent inflow, core, total	n/a	444 mm^2,0.16 m^2	444 mm^2,0.16 m^2	444 mm^2,0.29 m^2	at 40%, (7),(18),(26)

(c) One cubic-metre scenario	RackMount Case	Basic Tiled Cube	KC Single Packed	KC Double Packed	Eqns/Notes
p_1 Assembly mode	1x42U x 20 x 2U Each 2U: 12x12 PE	n=12 single packed face-centred-cubic	n=12 single packed face-centred-cubic	n=12 double packed bifurcated honeycomb	
p_2 Core Count , Power	2880 , 17W	1,728, 25W	1,728, 25W	3,059, 14W	(1),(12)
p_3 Required Power	49 kW	43 kW (2 pins)	43 kW (2 pins)	43 kW (1 pins)	pins per T-facet
p_4 Available Power	20	62 kW (144%)	62 kW (144%)	57 kW (134%)	(7),(18)
p_5 Core Size , System Size	n/a	80mm , 1.0m cube	80mm , 1.0m cube	80mm , 1.0m cube	(2)
p_6 System Volume	$1.2m^3$	$0.9m^3$	$0.9m^3$	$0.9m^3$	$(n \times d)^3$
p_7 Power Density	$40.8Kwh/m^3$	$48.8kW/m^3$	$48.8kW/m^3$	$48.4kW/m^3$	$p_3 \div p_6$
p_8 Core Bw, Tot Bisect. BW	n/a	960 Gbps , 11.5 Tbps	480 Gbps , 5.8 Tbps	960 Gbps , 63.8 Tbps	(10),(23)
p_9 T/H/G Bisection BW	n/a	(5.8, - , 5.8) Tbps	(2.9, - , 2.9) Tbps	(2.9, 58.1, 2.9) Tbps	
p_{10} Raw PetaFlops @50 GFLOPS/W	2.45 Pflops	2.2 PFlops	2.2 PFlops	2.1 PFlops	(11),(24)
p_{11} H-facets circ area	n/a	n/a	$11,833mm^2$	$11,833mm^2$	(28)
p_{12} Void Vol, per core, total	n/a	279 cm^3, 0.5 m^3	279 cm^3, 0.5 m^3	64 cm^3, 0.2 m^3	at 30% void ratio
p_{13} T-Vent inflow, core, total	n/a	284 mm^2,0.12 m^2	284 mm^2,0.12 m^2	284 mm^2,0.23 m^2	at 40%, (7),(18),(26)

(d) Future Exascale case	RackMount Case	Basic Tiled Cube	KC Single Packed	KC Double Packed	Eqns/Notes
p_1 Assembly mode	60x42Ux 20x2U Each 2U: 10x11 PE	n=41 single packed face-centred-cubic	n=41 single packed face-centred-cubic	n=41 double packed bifurcated honeycomb	
p_2 Core Count , Power	132,000 , 22W	68,921, 42W	68,921, 42W	132,921, 22W	(1),(12)
p_3 Required Power	2904 kW	2895 kW (8 pins)	2895 kW (8 pins)	2924 kW (5 pins)	pins per T-facet
p_4 Available Power	n/a	2,905 kW (100%)	2,905 kW (100%)	3,543 kW (121%)	(7),(18)
p_5 Core Size , System Size	$\approx 60x1.2m^3$	80mm , 3.3m cube	80mm , 3.3m cube	80mm , 3.3m cube	(2)
p_6 System Volume	$\approx 72m^3$	$35.3m^3$	$35.3m^3$	$35.3m^3$	$(n \times d)^3$
p_7 Power Density	$40.3Kwh/m^3$	$82.0kW/m^3$	$82.0kW/m^3$	$82.9kW/m^3$	$p_3 \div p_6$
p_8 Core Bw, Tot Bisect. BW	n/a	1920 Gbps , 269.0 Tbps	1920 Gbps , 269.0 Tbps	2560 Gbps , 1293.0 Tbps	(10),(23)
p_9 T/H/G Bisection BW	n/a	(134.5, - , 134.5) Tbps	(134.5, - , 134.5) Tbps	(134.5, 1024., 134.5) Tbps	
p_{10} Raw PetaFlops @400 GFLOPS/W	1161.6 Pflops	1071.0 PFlops	1071.0 PFlops	1082.0 PFlops	(11),(24)
p_{11} H-facets circ area	n/a	n/a	$11,833mm^2$	$11,833mm^2$	(28)
p_{12} Void Vol, per core, total	n/a	279 cm^3, 19.2 m^3	279 cm^3, 19.2 m^3	64 cm^3, 8.6 m^3	at 30% void ratio
p_{13} T-Vent inflow, core, total	n/a	284 mm^2,1.43 m^2	284 mm^2,1.43 m^2	284 mm^2,2.80 m^2	at 40%, (7),(18),(26)

bandwidth of over 2.5 Tbps per core. In this system the peak power consumption is of the order of 2,900 kWh region in all cases, necessitating a substantial cooling system infrastructure, though with a peak power density of under $90Kw/m^3$. However, the system's core array dimensions are equivalent to a 3.3 metre cube, or $\approx$ 35m^3, implying performance density of 66 PetaFLOP/m^3, although this excludes the extra cooling infrastructure present in all system assemblies - PTCA, cabinet or otherwise.

It can also be seen, taking the double-packed case for example, that the interior cooling network has an estimated volume of 8.6m^3 while the total inward flow vent area is almost 2.8m^2, which represents a significant potential flow capacity (be it air or liquid).

It is important to recognize that, in practice, feasible cooling mediums (air/liquid/etc) may have different properties (specific heat capacity, viscosity, etc.) and achieving equilibrium and consistent distribution of cooling with the effects of flows and mixing within the grid labyrinth therefore requires a much more complex analysis than we can present here. It should be possible to see however that as systems grow very large, the ratio of external cooling ports to internal volume will decrease, and at some point a limit will be reached for an individual modular array.

E. FUTURE SCALING TRENDS

As has been highlighted earlier, Koomey's law and the increasing levels of integration for IC fabrication suggest that at least in the near future there will continue to be considerable gains in power efficiency per GFLOPS and continuing increased levels of integration. The preceding evaluation cases show that core power and core size are primary factors in achieving manageable power density and useful levels of performance in a given form factor. The question then is, what levels of performance can be achieved in PTCA formats at the future point where power efficiency is reaching its immediately foreseeable limits in terms of traditional technology trends?

To partially answer this question an analysis is provided in Table 13, where the single- and double-packed arrays are considered at a range of K-Core dimensions from 40 mm up to 100 mm, and with power per core ranging from 5 W to 40 W, and selected design constraints corresponding to three scenarios. The first case, Tables 13(a) and (b), show a moderately high power density limit set at 150kW/m^3 and a currently contemporary 50 GFLOPS/Watt power efficiency rating. The second case assumes a Koomey's law doubling rate of 2.6 years over a period to 2030, with a 200kW/m^3 cooling limit. The third case, with the same cooling limit, takes a more conservative doubling rate of 3 years over a ten-year period to reflect a conservative view of Koomey's law over that period. These doubling rates approximate to a factor of 5 and 10 improvement respectively, whereas 'optimistic' Koomey's law trends suggest a ten-year increase of 16-fold.

A 40 mm core is quite small, allowing only around 19 $\times$ 19 mm per H-facet for circuitry at 80% usable area (Eq. (29)), whereas a 100 mm core is quite large and provides over 48 mm $\times$ 48 mm of circuit area per H-facet. Nonetheless, the T-facet for the $d = 40$ mm case would have an area of 178 mm^2 (Eq. (26)), which could accommodate over 40 power and IO pins at 100% utilization, a figure far higher than necessary for most arrays, meaning that there is capacity to accommodate both power and a limited IO channel connectivity on the same T-facet. The current trend toward mainstream 2 nm VLSI process technologies, advanced die stacking and integrated photonics, promises to make this quite feasible, while the much longer term view of unconventional computational devices beyond traditional silicon integration may present some unique opportunities for cores of this size and perhaps significantly smaller still.

Alongside power density estimates, Table 13 also presents the implied performance in terms of the theoretical maximum raw PetaFLOPS implied by the sustained power load available at each node. The base assumption of approximately 50 GFLOPS/watt is a figure already observed in a recent survey of CPU and GPU data [81]. The 'Green500' rankings cite the top ten systems as having GFLOPS/watt ranging from 56 to 72 GFLOPS/watt and averaging 64 GFLOPS/watt [82].

This analysis suggests that systems capable of over 100 raw PetaFLOPS/m^3 (or 0.1 ExaFLOPS/m^3) could be feasible at this first-order level of evaluation granularity, and with a variety of core sizes/counts, and on-board power loads.

V. FUTURE WORK ROAD-MAP

An important objective of the work reported here was to make an inroad into a design space that is sparsely recognized or explored, and to create a foundation for the work needed to move toward the next level of implementation feasibility and understanding. A number of 'road-map goals' must be reached in order to complete this picture, and each of these may represent significant pieces of research with wider benefits than solely the progression of PTCA technology. In particular the authors highlight the following areas that are important for further investigation:

- **Generic models for polyhedral capabilities**
  - A methodology to identify the appropriateness of any polyhedral shape is highly desirable.
  - Starting points include past work in literature [49] where a large number of polyhedra (over 50,000) have been iterated algorithmically.
  - This goal would permit a wide range of possible polyhedral shapes to be considered, allowing their individual advantages and limitations to be quantified, grouped and classified.
- **Air/Fluid flow models**
  - A framework to simulate air and fluid flow, in PTCA grids of any size and tiling structure, is needed.

TABLE 13. Power and performance design envelopes per cubic metre, for single and double packed scenarios (50-550 GFLOPS/Watt).

Single Packed Configurations

d(mm)	n	C_S	$H_{area\ mm^2}$
40	25	15,625	2,956
50	20	8,000	4,619
60	16	4,096	6,651
70	14	2,744	9,053
80	12	1,728	11,824
90	11	1,331	14,965
100	10	1,000	18,475

Core dimensions and Counts

5w	10w	15w	20w	25w	30w	35w	40w
2	4	4	6	6	8	10	10
2	2	4	4	6	6	8	8
2	2	4	4	4	6	6	6
2	2	2	4	4	4	6	6
2	2	2	4	4	4	4	6
2	2	2	4	4	4	4	6
2	2	2	2	4	4	4	4

T- Facet Power Pin Counts

Double Packed Configurations

d(mm)	n	C_D	$H_{area\ mm^2}$
40	25	29,449	2,956
50	20	14,859	4,619
60	16	7,471	6,651
70	14	4,941	9,053
80	12	3,059	11,824
90	11	2,331	14,965
100	10	1,729	18,475

Core dimensions and Counts

5w	10w	15w	20w	25w	30w	35w	40w
1	2	2	3	3	4	5	5
1	1	2	2	3	3	4	4
1	1	2	2	2	3	3	3
1	1	1	2	2	2	3	3
1	1	1	2	2	2	2	3
1	1	1	2	2	2	2	2
1	1	1	1	2	2	2	2

T- Facet Power Pin Counts

(a) Single Packed 50GFlops/Watt, 150kWh Cooling Limit

5w	10w	15w	20w	25w	30w	35w	40w
78	156	234	312	390	469	547	625
40	80	120	160	200	240	280	320
20	41	61	82	102	123	143	164
14	28	41	55	69	82	96	110
9	18	26	35	43	52	60	69
7	14	20	27	33	40	47	53
5	10	15	20	25	30	35	40

5w	10w	15w	20w	25w	30w	35w	40w
3.9	7.8	11.7	15.6	19.5	23.4	27.3	31.2
2.0	4.0	6.0	8.0	10.0	12.0	14.0	16.0
1.0	2.0	3.1	4.1	5.1	6.1	7.2	8.2
0.7	1.4	2.1	2.7	3.4	4.1	4.8	5.5
0.5	0.9	1.3	1.7	2.2	2.6	3.0	3.4
0.4	0.7	1.0	1.3	1.7	2.0	2.3	2.7
0.3	0.5	0.8	1.0	1.3	1.5	1.8	2.0

(b) Double Packed 50GFlops/Watt, 150kWh Cooling Limit

5w	10w	15w	20w	25w	30w	35w	40w
147	294	442	589	736	883	1031	1178
74	149	223	297	371	446	520	594
37	75	112	149	187	224	262	299
25	49	74	99	124	148	173	198
15	31	46	61	77	92	107	122
12	23	35	47	58	70	82	93
9	17	26	35	43	52	61	69

5w	10w	15w	20w	25w	30w	35w	40w
7.4	14.7	22.1	29.4	36.8	44.2	51.5	58.9
3.7	7.4	11.1	14.9	18.6	22.3	26.0	29.7
1.9	3.7	5.6	7.5	9.3	11.2	13.1	14.9
1.2	2.5	3.7	4.9	6.2	7.4	8.7	9.9
0.8	1.5	2.3	3.1	3.8	4.6	5.4	6.1
0.6	1.2	1.8	2.3	2.9	3.5	4.1	4.7
0.4	0.9	1.3	1.7	2.2	2.6	3.0	3.5

(c) Single Packed, 275GFlops/Watt, 200kWh Cooling Limit

5w	10w	15w	20w	25w	30w	35w	40w
147	294	442	589	736	883	1031	1178
74	149	223	297	371	446	520	594
37	75	112	149	187	224	262	299
25	49	74	99	124	148	173	198
15	31	46	61	77	92	107	122
12	23	35	47	58	70	82	93
9	17	26	35	43	52	61	69

5w	10w	15w	20w	25w	30w	35w	40w
7.4	14.7	22.1	29.4	36.8	44.2	51.5	58.9
3.7	7.4	11.1	14.9	18.6	22.3	26.0	29.7
1.9	3.7	5.6	7.5	9.3	11.2	13.1	14.9
1.2	2.5	3.7	4.9	6.2	7.4	8.7	9.9
0.8	1.5	2.3	3.1	3.8	4.6	5.4	6.1
0.6	1.2	1.8	2.3	2.9	3.5	4.1	4.7
0.4	0.9	1.3	1.7	2.2	2.6	3.0	3.5

(d) Double Packed, 275GFlops/Watt, 200kWh Cooling Limit

5w	10w	15w	20w	25w	30w	35w	40w
78	156	234	312	390	469	547	625
40	80	120	160	200	240	280	320
20	41	61	82	102	123	143	164
14	28	41	55	69	82	96	110
9	18	26	35	43	52	60	69
7	14	20	27	33	40	47	53
5	10	15	20	25	30	35	40

5w	10w	15w	20w	25w	30w	35w	40w
21	43	64	86	107	129	150	172
11	22	33	44	55	66	77	88
6	11	17	22	28	34	39	45
4	8	11	15	19	23	26	30
2	5	7	9	12	14	17	19
2	4	6	7	9	11	13	15
1	3	4	6	7	8	10	11

(e) Single Packed, 550GFlops/Watt, 200kWh Cooling Limit

5w	10w	15w	20w	25w	30w	35w	40w
78	156	234	312	390	469	547	625
40	80	120	160	200	240	280	320
20	41	61	82	102	123	143	164
14	28	41	55	69	82	96	110
9	18	26	35	43	52	60	69
7	14	20	27	33	40	47	53
5	10	15	20	25	30	35	40

5w	10w	15w	20w	25w	30w	35w	40w
21	43	64	86	107	129	150	172
11	22	33	44	55	66	77	88
6	11	17	22	28	34	39	45
4	8	11	15	19	23	26	30
2	5	7	9	12	14	17	19
2	4	6	7	9	11	13	15
1	3	4	6	7	8	10	11

(f) Double Packed, 550GFlops/Watt, 200kWh Cooling Limit

5w	10w	15w	20w	25w	30w	35w	40w
147	294	442	589	736	883	1031	1178
74	149	223	297	371	446	520	594
37	75	112	149	187	224	262	299
25	49	74	99	124	148	173	198
15	31	46	61	77	92	107	122
12	23	35	47	58	70	82	93
9	17	26	35	43	52	61	69

5w	10w	15w	20w	25w	30w	35w	40w
40	81	121	162	202	243	283	324
20	41	61	82	102	123	143	163
10	20	31	41	51	62	72	82
7	14	20	27	34	41	48	54
4	8	13	17	21	25	29	34
3	6	10	13	16	19	22	26
2	5	7	10	12	14	17	19

This table presents peak power in kW for single-packed array cases (a),(c),(e) and double packed array cases (b),(d),(f) for three GFLOPS/Watt and cooling limit scenarios. Colored zones represent air cooling (green), immersive liquid cooling (blue) or else cases currently considered outside a feasible cooling capability at that power density (red).

Informative work should be sought to guide this pursuit [50], [55], [56], [57], [58], [59].

- Models must take into account the properties of internal flow channel and void spaces, such as interior surface drag effects, turbulence, intermixing and so-on. An excellent example is given in [57].
- This goal would allow thermal properties of any PTCA system to be evaluated with given configurations for vent aperture, power load, grid size, etc.

- **Mapping and folding theories**
  - A methodology is needed for evaluating folding and mapping of logical topologies onto the physical PTCA topologies formed by any given polyhedral module type. This could be founded upon relevant work in the field including but not limited to work referenced in this paper [38], [39], [40], [41], [42].
  - This would permit automated validations, given some polyhedra type p_x, that it is feasible to map well known topologies onto structures with certain degrees of efficiency, with hop distances, latencies, traffic bandwidth, congestion factors and so-on being quantifiable, while also accounting for the optimization of thermal and power densities, hot-spots, and enhanced fault tolerance.
  - This goal may also permit certain unique or obscure topologies to be discovered as systems able to be implemented as PTCA systems.

- **Detailed HPC performance studies**
  - Work to model PTCA in direct comparison with particular HPC systems, with equitable workloads, would be very desirable.
  - This would require substantial work to simulate workloads on each system, extending earlier work on PTCA models using BookSim2 and Anynet modeling toolsets [83] as reported in [4] for example, whilst there are opportunities to develop power/energy/performance models through a number of possible strategies [84], [85], [86], [87], [88], [89], and perhaps including the widely exploited SST toolkit [90].
- **Enhanced resilience**
  - Designating some nodes in a grid as 'power reservoirs', to supplement local transient variations in peak power demand, is of interest and should be explored.
  - Modeling the fault-tolerance and MTBF of given system configurations is essential. Basic models of redundancy can be enhanced to provide design insights to develop strategies for managing fault behaviors.
  - The distinction between data processing component failures (e.g. CPU faults) versus power grid behavior demands each to seek its own methodology.
  - As mentioned in previous road-map goals, the ability to optimally map workloads, exploiting structures, neighborhoods and connectivities, in such a way to reduce criticality of fault effects is also then possible. Some excellent and relevant guidance is given in [43].
- **Push the limits of PTCA IO fabrics**
  - Extend the limits of wire-pin based IO facet interfaces, taking into account EMF and crosstalk effects and trade-offs to maximize facet bandwidths.
  - Explore the potential for economically viable silicon photonics solutions to deliver much more IO bandwidth and channel multiplicity. This is a very active field, and we can expect step changes in capabilities versus component cost within this decade [63], [64], [65], [66], [67], [68], [70], [72], [91], [92].
- **Understand physical connector effects**
  - Including effects of cooling fluids on electrical and optical connectors, also exploring other connector concepts which might increase grid capabilities.
  - Dielectric breakdown voltages of coolants at short (millimetre scale) ranges and impedance effects of high frequency multi-GigaHertz IO signals of closely grouped pins immersed in fluids with such properties will need to be evaluated in detail, including practical test-bench experimentation.
  - Very little work has been done in this area, but interest in this topic is emerging [93], [94]. Limitations need to be better understood and new solutions engineered.

Importantly, the insights gained from any of these road-map goals are not only to model and understand systems, but also to highlight areas for specific targeted technology developments. No doubt, substantial effort is implied to progress in any of these areas, and therefore a full-scale development of PTCA systems is still a major research challenge.

More broadly, the work may inform the evolution of other novel system architecture solutions that may succeed current mainstream technologies. There is much still to be learned.

VI. CONCLUSION

PTCA is a novel and as yet not widely investigated paradigm for composing systems which scale in three dimensions, based upon the tiling of polyhedra . It offers unique capabilities for future HPC systems that are not yet fully explored, and a design space that deserves much more attention in theoretical, simulation and engineering terms.

Ultimately there are a huge number of possible PTCA implementations to consider if a thorough mapping of the design space were to be attempted. Simulations at large scales are time-consuming and resource-intensive [1], while building large-scale prototypes without exploring the fundamentals first would be premature. Prototypes such as the currently in-progress 'K1 array' project do allow subtle trade-offs to be exposed as well as the hard engineering challenges that may come with such an unusual paradigm.

Complementing these approaches, as presented here, is a mathematical approach based on the geometry, topology and characteristics of hypothetical polyhedral candidates, a technique easily adaptable to any other polyhedron. A key goal was to use this approach to test the four key questions around feasibility posed at the start of this paper.

Returning to the essential needs of unconventional future HPC systems as recently established by Becker et al. [14], it is arguable that PTCA has also met all of those cited requirements to an extent that shows PTCA deserves further research with many new questions yet to be explored.

Current work on the K1 prototype array is progressing. This is in many ways a modest implementation, based upon a $3 \times 3 \times 3$ double packed cube. Though limited in scale, this will serve as a learning exercise in understanding engineering problems alongside theoretical projections, and will demonstrate its practical operation with sample workloads. Moving beyond that initial goal will require attention to engineering detail and a step-change in the sophistication of components used in module interfaces for power and IO. However it is entirely feasible to build an array of the order of 1000 cores with existing technology. The main obstacle at this moment is availability of funding rather than a technical road-block. The authors look forward to expanding this work if that can be addressed.

This paper has presented an exposition of a very different system assembly paradigm to the traditional approaches, and PTCA is still a barely explored concept. The inter-relatedness of many of the highlighted concepts mean this is an area with a great deal yet to be understood. Nonetheless, with

the desire for ever more complex data and AI systems, the need for increasingly complex processing structures and the consequences of ever higher performance densities at the chip level, exploring alternative design spaces must be a worthwhile endeavor, and PTCA offers ample opportunity to do so.

AUTHOR CONTRIBUTIONS

Jim Austin has developed the original idea, and patent for ''ball computing'' [2], whereby hexagonal tiles can form into structures such as truncated octahedra. Chris Crispin-Bailey has identified the mathematical models for deriving properties of the presented polyhedral modules from primary module parameters, and acted as a primary author for this article. Pakon Thuphairo has contributed to the work on power grid modeling [4]. Anthony Moulds has developed hardware implementations of Hex-Tile and K1 array polyhedral prototypes as part of a design and build project with Dr. Chris Crispin-Bailey. Steven Wright has contributed to the knowledge and insights into HPC systems, and in particular those that target 3-D logical topologies for computational problems.

CONFLICTS OF INTEREST

Prof. Jim Austin (retired) holds U.K. Patent (No. GB2529617), relating to some aspects of tileable computing modules. There is no known commercial interest at the time of publication.

REFERENCES

[1] P. Thuphairo, C. Bailey, A. Moulds, and J. Austin, ''Investigating novel 3D modular schemes for large array topologies: Power modeling and prototype feasibility,'' in *Proc. 25th Euromicro Conf. Digit. Syst. Design (DSD)*, Aug. 2022, pp. 268–275.

[2] J. Austin, ''Computing devices,'' U.K. Patent GB 2 529 617, Aug. 5, 2014.

[3] A. M. Kamali Sarvestani, C. Bailey, and J. Austin, ''Performance analysis of a 3D wireless massively parallel computer,'' *J. Sensor Actuator Netw.*, vol. 7, no. 2, p. 18, Apr. 2018.

[4] P. Thuphairo, ''Modelling and simulation for power distribution grids of 3D tiled computing arrays,'' Ph.D. dissertation, Dept. Comput. Sci., University of York, York, U.K., 2023.

[5] A. M. K. Sarvestani, ''Evaluating techniques for wireless interconnected 3D processor arrays,'' Ph.D. dissertation, Dept. Comput. Sci., University of York, York, U.K., 2013.

[6] L. A. Plana, J. Garside, J. Heathcote, J. Pepper, S. Temple, S. Davidson, M. Luján, and S. Furber, ''SpiNNlink: FPGA-based interconnect for the million-core SpiNNaker system,'' *IEEE Access*, vol. 8, pp. 84918–84928, 2020.

[7] N. Jouppi, G. Kurian, S. Li, P. Ma, R. Nagarajan, L. Nai, N. Patil, S. Subramanian, A. Swing, B. Towles, C. Young, X. Zhou, Z. Zhou, and D. A. Patterson, ''TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings,'' in *Proc. 50th Annu. Int. Symp. Comput. Archit.*, Jun. 2023, pp. 1–14.

[8] Y. Sverdlik. (2018). *Google Brings Liquid Cooling to Data Centers To Cool Latest AI Chips*. [Online]. Available: https://www.datacenterknowledge.com/google-alphabet/google-brings-liquid-cooling-data-centers-cool-latest-ai-chips

[9] I. Ohmura, G. Morimoto, Y. Ohno, A. Hasegawa, and M. Taiji, ''MDGRAPE-4: A special-purpose computer system for molecular dynamics simulations,'' *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 372, no. 2021, Aug. 2014, Art. no. 20130387.

[10] P. Hamm, ''Toward an FPGA-based dedicated computer for molecular dynamics simulations,'' *J. Chem. Phys.*, vol. 162, no. 5, pp. 1–14, Feb. 2025.

[11] R. H. Larson, J. K. Salmon, R. O. Dror, M. M. Deneroff, C. Young, J. P. Grossman, Y. Shan, J. L. Klepeis, and D. E. Shaw, ''High-throughput pairwise point interactions in Anton, a specialized machine for molecular dynamics simulation,'' in *Proc. IEEE 14th Int. Symp. High Perform. Comput. Archit.*, Feb. 2008, pp. 331–342.

[12] R. O. Dror, J. P. Grossman, K. M. Mackenzie, B. Towles, E. Chow, J. K. Salmon, C. Young, J. A. Bank, B. Batson, M. M. Deneroff, J. S. Kuskin, R. H. Larson, M. A. Moraes, and D. E. Shaw, ''Exploiting 162-nanosecond end-to-end communication latency on Anton,'' in *SC : Proc. ACM/IEEE Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Nov. 2010, pp. 1–12.

[13] P. Ruch, T. Brunschwiler, W. Escher, S. Paredes, and B. Michel, ''Toward five-dimensional scaling: How density improves efficiency in future computers,'' *IBM J. Res. Develop.*, vol. 55, no. 5, pp. 15:1–15:13, Sep. 2011.

[14] T. Becker, R. Haas, J. Schemmel, S. Furber, and S. Dolas, ''Unconventional hpc architectures,'' Zenodo, Apr. 2022, doi: 10.5281/zenodo.6470840.

[15] A. H. Khalaj and S. K. Halgamuge, ''A review on efficient thermal management of air- and liquid-cooled data centers: From chip to the cooling system,'' *Appl. Energy*, vol. 205, pp. 1165–1188, Nov. 2017.

[16] A. G. D. Rowley, C. Brenninkmeijer, S. Davidson, D. Fellows, A. Gait, D. R. Lester, L. A. Plana, O. Rhodes, A. B. Stokes, and S. B. Furber, ''SpiNNTools: The execution engine for the SpiNNaker platform,'' *Frontiers Neurosci.*, vol. 13, p. 231, Mar. 2019.

[17] S. Furber and P. Bogdan, *SpiNNaker: A Spiking Neural Network Architecture*. Boston, MA, USA: Now, Mar. 2020.

[18] S. J. van Albada, A. G. Rowley, J. Senk, M. Hopkins, M. Schmidt, A. B. Stokes, D. R. Lester, M. Diesmann, and S. B. Furber, ''Performance comparison of the digital neuromorphic hardware SpiNNaker and the neural network simulation software NEST for a full-scale cortical microcircuit model,'' *Frontiers Neurosci.*, vol. 12, p. 291, May 2018.

[19] E. Painkras, L. A. Plana, J. Garside, S. Temple, S. Davidson, J. Pepper, D. Clark, C. Patterson, and S. Furber, ''SpiNNaker: A multi-core system-on-chip for massively-parallel neural net simulation,'' in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2012, pp. 1–4.

[20] A. Heydari, A. R. Gharaibeh, M. Tradat, Q. Soud, Y. Manaserh, V. Radmard, B. Eslami, J. Rodriguez, and B. Sammakia, ''Experimental evaluation of direct-to-chip cold plate liquid cooling for high-heat-density data centers,'' *Appl. Thermal Eng.*, vol. 239, Feb. 2024, Art. no. 122122.

[21] K. W. Yan, P.-Y. Lin, and S.-L. Kuo, ''Thermal challenges for HPC 3DFabricTM packages and systems,'' in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Mar. 2022, pp. 1–4.

[22] C. Conficoni, A. Bartolini, A. Tilli, C. Cavazzoni, and L. Benini, ''HPC cooling: A flexible modeling tool for effective design and management,'' *IEEE Trans. Sustain. Comput.*, vol. 6, no. 3, pp. 441–455, Jul. 2021.

[23] A. Banerjee, T. Mukherjee, G. Varsamopoulos, and S. K. S. Gupta, ''Cooling-aware and thermal-aware workload placement for green HPC data centers,'' in *Proc. Int. Conf. Green Comput.*, Aug. 2010, pp. 245–256.

[24] G. S. Roopan and M. D. Vijayakumar, ''A comprehensive review of CPU cooling systems: Innovations, efficiency, and future directions,'' in *Proc. Int. Conf. Adv. Additive Manuf. Technol.*, 2024, pp. 375–380.

[25] I.-H. Chung, T. N. Sainath, B. Ramabhadran, M. Picheny, J. Gunnels, V. Austel, U. Chauhari, and B. Kingsbury, ''Parallel deep neural network training for big data on blue Gene/Q,'' *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 6, pp. 1703–1714, Jun. 2017.

[26] P. Vivet et al., ''A 220GOPS 96-core processor with 6 chiplets 3D-stacked on an active interposer offering 0.6ns/mm latency, 3Tb/s/mm^2 inter-chiplet interconnects and 156 mW/mm^2 @ 82%-peak-efficiency DC–DC converters,'' in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 46–48.

[27] T. S. Komatsu, N. Okimoto, Y. M. Koyama, Y. Hirano, G. Morimoto, Y. Ohno, and M. Taiji, ''Drug binding dynamics of the dimeric SARS-CoV-2 main protease, determined by molecular dynamics simulation,'' *Sci. Rep.*, vol. 10, no. 1, p. 16986, Oct. 2020.

[28] A. R. Rao, G. A. Cecchi, and M. Magnasco, ''High performance computing environment for multidimensional image analysis,'' *BMC Cell Biol.*, vol. 8, no. 1, pp. 1–9, Jul. 2007.

[29] B. G. Fitch, R. S. Germain, M. Mendell, J. Pitera, M. Pitman, A. Rayshubskiy, Y. Sham, F. Suits, W. Swope, T. J. C. Ward, Y. Zhestkov, and R. Zhou, ''Blue matter, an application framework for molecular simulation on blue gene,'' *J. Parallel Distrib. Comput.*, vol. 63, nos. 7–8, pp. 759–773, Jul. 2003.

[30] H. Yoshida, Y. Wu, W. Cai, and B. Brett, "Scalable, high-performance 3D imaging software platform: System architecture and application to virtual colonoscopy," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2012, pp. 3994–3997.

[31] M. Taiji, T. Narumi, Y. Ohno, and A. Konagaya, "MDGRAPE-3: A petaflops special-purpose computer system for molecular dynamics simulations," in *Proc. Parallel Comput.*, 2004, pp. 669–676. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927545204800832

[32] D. E. Shaw et al., "Anton 3: Twenty microseconds of molecular dynamics simulation before lunch," in *Proc. SC21: Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Nov. 2021, pp. 1–11.

[33] F. Allen et al., "Blue gene: A vision for protein science using a petaflop supercomputer," *IBM Syst. J.*, vol. 40, no. 2, pp. 310–327, 2001.

[34] A. Probst, T. Knopp, C. Grabe, and J. Jägersküpper, "HPC requirements of high-fidelity flow simulations for aerodynamic applications," in *Proc. Eur. Conf. Parallel Process.*, 2020, pp. 375–387.

[35] J. Sheng, C. Yang, A. Sanaullah, M. Papamichael, A. Caulfield, and M. C. Herbordt, "HPC on FPGA clouds: 3D FFTs and implications for molecular dynamics," in *Proc. 27th Int. Conf. Field Program. Log. Appl. (FPL)*, Sep. 2017, pp. 1–4.

[36] Z. Zhou, Z. Li, W. Zhou, N. Chi, J. Zhang, and Q. Dai, "Resource-saving and high-robustness image sensing based on binary optical computing," *Laser Photon. Rev.*, vol. 19, no. 7, Apr. 2025, Art. no. 2400936.

[37] J. Nguyen, J. S. Pezaris, G. A. Pratt, and S. Ward, "Three-dimensional network topologies," in *Proc. 1st Int. Workshop Parallel Comput. Routing Commun.*, Washington, DC, USA. Cham, Switzerland: Springer, 1994, pp. 101–115.

[38] A. A. Ravankar and S. G. Sedukhin, "Mesh-of-Tori: A novel interconnection network for frontal plane cellular processors," in *Proc. 1st Int. Conf. Netw. Comput.*, Nov. 2010, pp. 281–284.

[39] P.-L. Lai and C.-H. Tsai, "Embedding of tori and grids into twisted cubes," *Theor. Comput. Sci.*, vol. 411, nos. 40–42, pp. 3763–3773, Sep. 2010.

[40] P. J. Roig, S. Alcaraz, K. Gilly, C. Bernad, and C. Juiz, "Arithmetic modeling of k-ary n-cubes and toroidal k-ary grids," *J. Phys., Conf. Ser.*, vol. 2701, no. 1, Feb. 2024, Art. no. 012036.

[41] H. Yu, I.-H. Chung, and J. Moreira, "Topology mapping for blue Gene/L supercomputer," in *Proc. ACM/IEEE SC Conf. (SC)*, Nov. 2006, pp. 116–es.

[42] T. Hoefler and M. Snir, "Generic topology mapping strategies for large-scale parallel architectures," in *Proc. Int. Conf. Supercomputing*, May 2011, pp. 75–84.

[43] I. Takanami, *Self-restructuring in Fault Tolerant Architecture: Processor Arrays with Spares*. Cham, Switzerland: Springer, 2025.

[44] W. Thomson, "On the division of space with minimum partitional area," *Acta Mathematica*, vol. 11, no. 1, pp. 121–134, 1887.

[45] D. Weaire and R. Phelan, "A counter-example to Kelvin's conjecture on minimal surfaces," *Phil. Mag. Lett.*, vol. 69, no. 2, pp. 107–110, Feb. 1994.

[46] R. Strand, B. Nagy, and G. Borgefors, "Digital distance functions on three-dimensional grids," *Theor. Comput. Sci.*, vol. 412, no. 15, pp. 1350–1363, Mar. 2011.

[47] S. Torquato and Y. Jiao, "Dense packings of the platonic and Archimedean solids," *Nature*, vol. 460, no. 7257, pp. 876–879, Aug. 2009.

[48] R. Gabbrielli, Y. Jiao, and S. Torquato, "Families of tessellations of space by elementary polyhedra via retessellations of face-centered-cubic and related tilings," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 86, no. 4, Oct. 2012, Art. no. 041141.

[49] E. R. Chen, D. Klotsa, M. Engel, P. F. Damasceno, and S. C. Glotzer, "Complexity in surfaces of densest packings for families of polyhedra," *Phys. Rev. X*, vol. 4, no. 1, Feb. 2014, Art. no. 011024.

[50] S. Wang, Y. Yin, C. Hu, and P. Rezai, "3D integrated circuit cooling with microfluidics," *Micromachines*, vol. 9, no. 6, p. 287, Jun. 2018.

[51] S. Torii and H. Ishikawa, "Zettascaler: Liquid immersion cooling many-core based supercomputer," *CANDAR2017 Keynote*, vol. 3, 2017.

[52] F. M. Naduvilakath-Mohammed, R. Jenkins, G. Byrne, and A. J. Robinson, "Closed loop liquid cooling of high-powered CPUs: A case study on cooling performance and energy optimization," *Case Stud. Thermal Eng.*, vol. 50, Apr. 2023, Art. no. 103472.

[53] R. Wang, J. Qian, T. Wei, and H. Huang, "Integrated closed cooling system for high-power chips," *Case Stud. Thermal Eng.*, vol. 26, Aug. 2021, Art. no. 100954.

[54] P. Taddeo, J. Romaní, J. Summers, J. Gustafsson, I. Martorell, and J. Salom, "Experimental and numerical analysis of the thermal behaviour of a single-phase immersion-cooled data centre," *Appl. Thermal Eng.*, vol. 234, Nov. 2023, Art. no. 121260.

[55] T. Koch, "Projection-based resolved interface 1D-3D mixed-dimension method for embedded tubular network systems," *Comput. Math. Appl.*, vol. 109, pp. 15–29, Jun. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089812212200027X

[56] Y. Dai, R. Zhang, Z. Qin, K. Liu, C. Liu, and J. Zhao, "Research on the thermal performance and stability of three-dimensional array pulsating heat pipe for active/passive coupled thermal management application," *Appl. Thermal Eng.*, vol. 245, May 2024, Art. no. 122793. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1359431124004617

[57] A. K. Kareem, A. H. Turki, and A. M. Mohsen, "Optimization of turbulent flow heat transfer in a 3D cubic shell heat exchanger using non-mixture multiphase nanofluids," *J. Eng. Res.*, Oct. 2024, doi: 10.1016/j.jer.2024.10.013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2307187724002712

[58] D. V. Sheth and S. K. Saha, "Numerical study of thermal management of data centre using porous medium approach," *J. Building Eng.*, vol. 22, pp. 200–215, Mar. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235271021830812X

[59] J. Choi, Y. Kim, A. Sivasubramaniam, J. Srebric, Q. Wang, and J. Lee, "Modeling and managing thermal profiles of rack-mounted servers with ThermoStat," in *Proc. IEEE 13th Int. Symp. High Perform. Comput. Archit.*, Jun. 2007, pp. 205–215.

[60] C. Jang, S. Park, B. Infantolino, L. Lehman, R. Morgan, and D. Sengupta, "Failure analysis of contact probe pins for SnPb and sn applications," *Microelectron. Rel.*, vol. 48, no. 6, pp. 942–947, Jun. 2008.

[61] Q. Ping, W. Youwei, C. Wenhua, W. Zhe, and W. Yujie, "Contact reliability design modeling for wire spring-hole electrical connectors," *Microelectron. Rel.*, vol. 148, Sep. 2023, Art. no. 115182.

[62] L. Chang, M. Dijkstra, N. Ismail, M. Pollnau, R. M. de Ridder, K. Wörhoff, V. Subramaniam, and J. S. Kanger, "Waveguide-coupled micro-ball lens array suitable for mass fabrication," *Opt. Exp.*, vol. 23, no. 17, pp. 22414–22423, 2015.

[63] B. Ciftcioglu, R. Berman, S. Wang, J. Hu, I. Savidis, M. Jain, D. Moore, M. Huang, E. G. Friedman, G. Wicks, and H. Wu, "3-D integrated heterogeneous intra-chip free-space optical interconnect," *Opt. Exp.*, vol. 20, no. 4, pp. 4331–4345, 2012.

[64] J. A. Kash et al., "Chip-to-chip optical interconnects," in *Proc. Opt. Fiber Commun. Conf. Nat. Fiber Optic Eng. Conf.*, 2006, pp. 1–6.

[65] M. Blaicher, M. R. Billah, J. Kemal, T. Hoose, P. Marin-Palomo, A. Hofmann, Y. Kutuvantavida, C. Kieninger, P.-I. Dietrich, M. Lauermann, S. Wolf, U. Troppenz, M. Moehrle, F. Merget, S. Skacel, J. Witzens, S. Randel, W. Freude, and C. Koos, "Hybrid multi-chip assembly of optical communication engines by in situ 3D nano-lithography," *Light, Sci. Appl.*, vol. 9, no. 1, p. 71, Apr. 2020.

[66] H. Mekawey, M. Elsayed, Y. Ismail, and M. A. Swillam, "Optical interconnects finally seeing the light in silicon photonics: Past the hype," *Nanomaterials*, vol. 12, no. 3, p. 485, Jan. 2022.

[67] S. Bernabé, C. Kopp, M. Volpert, J. Harduin, J.-M. Fédéli, and H. Ribot, "Chip-to-chip optical interconnections between stacked self-aligned SOI photonic chips," *Opt. Exp.*, vol. 20, no. 7, pp. 7886–7894, 2012.

[68] A. Rizzo, A. Novick, V. Gopal, B. Y. Kim, X. Ji, S. Daudlin, Y. Okawachi, Q. Cheng, M. Lipson, A. L. Gaeta, and K. Bergman, "Massively scalable Kerr comb-driven silicon photonic link," *Nature Photon.*, vol. 17, no. 9, pp. 781–790, Sep. 2023.

[69] P.-J. Lu, M.-C. Lai, and J.-S. Chang, "A survey of high-performance interconnection networks in high-performance computer systems," *Electronics*, vol. 11, no. 9, p. 1369, Apr. 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/9/1369

[70] K. Takemura, D. Ohshima, A. Noriki, D. Okamoto, A. Ukita, J. Ushida, M. Tokushima, T. Shimizu, I. Ogura, D. Shimura, T. Aoki, T. Amano, and T. Nakamura, "Silicon-photonics-embedded interposers as co-packaged optics platform," *Trans. Jpn. Inst. Electron. Packag.*, vol. 15, no. 4, pp. 1–13, 2022.

[71] I. A. Young, E. Mohammed, J. T. Liao, A. M. Kern, S. Palermo, B. A. Block, M. R. Reshotko, and P. L. Chang, "Optical I/O technology for tera-scale computing," *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 235–248, Jan. 2010.

[72] T. Thorn and M. Vallo. (2017). *Leveraging Optical Chip-to-Chip Connectivity to Unleash the Complete Potential of AI*. [Online]. Available: https://www.yolegroup.com/player-interviews/leveraging-optical-chip-to-chip-connectivity-for-unleashing-the-complete-potential-of-ai-an-interview-with-ayar-labs/

[73] S. Han, T. Kim, J. Kim, and J. Kim, "A 10 Gbps SerDes for wireless chip-to-chip communication," in *Proc. Int. SoC Design Conf. (ISOCC)*, Nov. 2015, pp. 17–18.

[74] A. A. Suhani S. H., N. Prasad S., and K. S. S. Reddy, "A 20 Gb/s latency optimized SerDes transmitter for data centre applications," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT)*, Jul. 2020, pp. 1–4.

[75] H. Okuhara, A. Elnaqib, D. Rossi, A. Di Mauro, P. Mayer, P. Palestri, and L. Benini, "An energy-efficient low-voltage swing transceiver for mW-range IoT end-nodes," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Oct. 2020, pp. 1–5.

[76] J. Sundaram, S. Gopal, T. P. Thomas, E. Burton, and E. Ramirez, "A reconfigurable asynchronous SERDES for heterogenous chiplet interconnects," in *Proc. 22nd Int. Symp. Quality Electron. Design (ISQED)*, Apr. 2021, pp. 542–546.

[77] W. J. Turner, J. W. Poulton, J. M. Wilson, X. Chen, S. G. Tell, M. Fojtik, T. H. Greer, B. Zimmer, S. Song, N. Nedovic, S. S. Kudva, S. R. Sudhakaran, R. Bashirullah, W. Zhao, W. J. Dally, and C. T. Gray, "Ground-referenced signaling for intra-chip and short-reach chip-to-chip interconnects," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2018, pp. 1–8.

[78] V. Sriboonlue, Y. Jeon, G. R. Luevano, C. Ferguson, and E. Ochoa, "Comprehensive socket characterization and correlation for high-speed interface testing system," in *Proc. IEEE 74th Electron. Compon. Technol. Conf. (ECTC)*, May 2024, pp. 1768–1772.

[79] Z. Yang, Y. Gao, S. Li, C. Sun, and Y. Zhao, "Research on signal integrity in high-speed interconnection channel based on SIwave," in *Proc. IEEE 3rd Int. Conf. Circuits Syst. (ICCS)*, Oct. 2021, pp. 78–82.

[80] M. R. S. Katebzadeh, P. Costa, and B. Grot, "Evaluation of an InfiniBand switch: Choose latency or bandwidth, but not both," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Aug. 2020, pp. 180–191.

[81] Y. Sun, N. B. Agostini, S. Dong, and D. Kaeli, "Summarizing CPU and GPU design trends with product data," 2019, *arXiv:1911.11313*.

[82] (2023). *Green500 November 2023 Report*. Accessed: Jan. 3, 2024. [Online]. Available: https://www.top500.org/lists/green500/2023/11/

[83] N. Jiang, D. U. Becker, G. Michelogiannakis, J. Balfour, B. Towles, D. E. Shaw, J. Kim, and W. J. Dally, "A detailed and flexible cycle-accurate network-on-chip simulator," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Apr. 2013, pp. 86–96.

[84] M. Y. Teh, Z. Wu, M. Glick, S. Rumley, M. Ghobadi, and K. Bergman, "Performance trade-offs in reconfigurable networks for HPC," *J. Opt. Commun. Netw.*, vol. 14, no. 6, pp. 454–468, Jun. 2022.

[85] K. O'brien, I. Pietri, R. Reddy, A. Lastovetsky, and R. Sakellariou, "A survey of power and energy predictive models in HPC systems and applications," *ACM Comput. Surveys*, vol. 50, no. 3, pp. 1–38, May 2018.

[86] S. Farrell et al., "MLPerf HPC: A holistic benchmark suite for scientific machine learning on HPC systems," in *Proc. IEEE/ACM Workshop Mach. Learn. High Perform. Comput. Environ. (MLHPC)*, Nov. 2021, pp. 33–45.

[87] Z. Zhou, J. Sun, and G. Sun, "Automated HPC workload generation combining statistical modeling and autoregressive analysis," in *Proc. Int. Symp. Benchmarking*, 2024, pp. 153–170.

[88] M. Mubarak, C. D. Carothers, R. B. Ross, and P. Carns, "Enabling parallel simulation of large-scale HPC network systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 1, pp. 87–100, Jan. 2017.

[89] S. Böhm and C. Engelmann, "XSim: The extreme-scale simulator," in *Proc. Int. Conf. High Perform. Comput. Simul.*, Jul. 2011, pp. 280–286.

[90] A. Rodrigues, K. S. Hemmert, B. Barrett, C. Kersey, R. A. Oldfield, M. Weston, R. Risen, J. Cook, P. Rosenfeld, E. Cooper-Balis, and B. Jacob, "The structural simulation toolkit," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 38, no. 4, pp. 37–42, 2011.

[91] L. Virot, "Ultrafast on-chip germanium photodiode," *Nature Photon.*, vol. 15, no. 12, pp. 868–869, Dec. 2021.

[92] A. K. Kodi and A. Louri, "Optisim: A system simulation methodology for optically interconnected HPC systems," *IEEE Micro*, vol. 28, no. 5, pp. 22–36, Sep. 2008.

[93] Y.-S. Lo, J. Wu, L. Guo, Y. Huang, T. Xu, X. Mao, C. Chen, J. Zhang, J. Ho, and A. Liang, "A research on high speed signal integrity design optimal for immersion cooled server," in *Proc. 23rd IEEE Intersociety Conf. Thermal Thermomechanical Phenomena Electron. Syst. (ITherm)*, May 2024, pp. 1–8.

[94] A. Raniwala, "Bringing 2-phase immersion cooling to hyperscale cloud," in *Proc. Opt. Fiber Commun. Conf. Exhib. (OFC)*, Mar. 2022, pp. 1–3.

CHRIS CRISPIN-BAILEY received the Ph.D. degree in 1996. He is a Senior Lecturer in microelectronics and computer systems. He joined the Advanced Computer Architectures Group, University of York, in 1999, exploring low power systems design. He is currently a member of the Real-Time and Distributed Systems Group. He has worked on a 65 nm CMOS ASIC Fabrication Project (low power data compression), alongside miniature motion sensors for animal kinematics studies and novel CPU architectures and systems.

PAKON THUPHAIRO received the B.Eng. degree (Hons.) in computer engineering from Siam University, in 2005, the M.Eng. degree in computer engineering from Chulalongkorn University, in 2009, and the Ph.D. degree in computer science from the University of York, in 2024. He is currently a Lecturer with the Department of Computer Engineering, Faculty of Engineering, Rajamangala University of Technology Rattanakosin. His current main research interests include FPGA applications, digital logic circuits, and interconnection networks.

STEVEN WRIGHT is a Senior Lecturer in computer science with the University of York. He is a member of the Real-Time and Distributed Systems Research Group. He has collaborated widely with colleagues from national laboratories, universities, and industry, including U.S. Department of Energy, U.K. Atomic Energy Authority, Rolls-Royce, Intel, NVIDIA, ARM, and IBM. His research is broadly focused on high performance computing, in particular looking at the performance, programmability and energy efficiency of supercomputers and the scientific computing applications running on them.

ANTHONY MOULDS is a highly experienced Electronics Engineer and holds the post of a Senior Experimental Officer with the Department of Computer Science, University of York. He has significant experience in the development of ultra-low-power systems and particular expertise in the design of intelligent miniaturized power-efficient hardware and devices. His work in support of tileable computing modules, including the HexTile project and K1 array prototype, have been foundational to the development of polyhedral computing modules.

JIM AUSTIN is a Professor (retired) of neural computing. He was also a CEO of Cybula Ltd., a successful U.K. technology research and development consultancy. He manages a large museum collection of supercomputer systems and other computers of historical importance "The Jim Austin Computer Collection." He has interests in massively parallel array processing, scalability, pattern matching architectures, and data analytics. In 2012, he received the Times Higher Education Award for the Outstanding Research Team of the Year for work on distributed air-craft maintenance data-analytics.