Investigating Novel 3D Modular Schemes for Large Array Topologies:

Power Modeling and Prototype Feasibility.

Pakon Thuphairo
Dept. of Computer Science
University of York
York, United Kingdom
pt795@york.ac.uk

Christopher Bailey
Dept. of Computer Science
University of York
York, United Kingdom
chrisb@cs.york.ac.uk

Anthony Moulds
Dept. of Computer Science
University of York
York, United Kingdom
anthony.moulds@york.ac.uk

Jim Austin
Dept. of Computer Science
University of York (retired)
York, United Kingdom
jim.austin@york.ac.uk

***Abstract*— This paper presents the Tiled Computing Array (TCA), a simple, uniform, 3D-mesh packaging at inter-board level, for massively parallel computers. In particular, the power modelling and practical feasibility of the system is examined. TCA eliminates the need for hierarchical rackmount-structures and introduces short and immediate data channels in multiple physical orientations, allowing a more direct physical mapping of 3D computational topology to real hardware. A dedicated simulation platform has been developed, and an engineered prototype demonstrator has been built. This paper explores the feasibility of the TCA concept for current hardware technologies and systems, evaluates power modeling and validation, and highlights some of the novel design challenges associated with such a system. Evaluations of physical scalability toward large-scale systems are reported, showing that TCA is a promising approach.**

Keywords— computing array, interconnection network, massively parallel computers, scalability, simulation

I. Introduction

The complexity of hardware structures in the building of parallel computers is significant in terms of the effort of composing complete functional systems, starting with the processor chip as a fundamental building block, alongside memory devices, SSDs (solid-state drives) and communication ICs (integrated circuits), and power conditioning components. There are inherent complications in addressing this via board-level design, rackmount, and modular system hierarchies, and these physical demands create topological compromises between the logical processing structure and the physical equivalent. These are evident in terms of wiring constraints, power delivery and heat dissipation, and in terms of computational density of such systems.

In this paper, we aim to investigate a completely different approach, aiming to address such difficulties with a completely different structural paradigm, based upon fundamental building blocks, referred to as tiles, or 'hex-tiles'. Tiles are therefore modules containing one or more chips, perhaps ultimately embodied as an adaptation of existing well established IC packaging technology encapsulating with a single SoC die or perhaps a multi-chip module (MCM). Initial prototypes are necessarily less sophisticated and rely upon PCB level IC integration to create tile modules, an order of magnitude larger in scale, but capable of demonstrating concepts and principles.

Tiles as fundamental building blocks are capable of being tessellated in multiple ways. Due to a novel angled edge-interface arrangement, a group of eight tiles may be composed into a 3D structure which we equate to a 'ball'. Balls may then be coupled to each other to build larger systems, also extending directly in three dimensions as uniform arrays. Hex-tiles directly connect power and IO to one another, completing the power and data grids without additional circuit boards, racks, etc. This of course results in power delivery challenges, IO connectivity, and latency issues in this new model. In order to extend the knowledge of such systems and assess their viability, we present a conceptual model, a prototype, and a simulation tool which is used to investigate how these electrical constraints impact upon the scalability and feasibility of the system.

II. Motivation

Current state of the art massively parallel computing systems relies heavily upon the well-established technologies of back-plane, rackmount, and server cabinet infrastructures, along with the associated power bus architectures and interconnection strategies. Obviously, most of the systems are comprised of the supporting infrastructure, and relatively small parts of the system are the actual CPU, memory bank, SSD or other resources. In effect, the desired high-density collection of processing elements is forced to map onto a variety of physical inter-board level construction constraints, many if not all of which then impact upon other critical factors such as interconnection length, cooling strategies, granularity of local versus inter-module communications, and so-on.

The motivation for the tiled computing array (TCA) stems from this observation, and the question *'how can we interface maximum processing elements with minimal infrastructure and constraints, whilst also facilitating effective air or liquid cooling?'*. The TCA concept eliminates the need for rackmount architectures and permits a more direct physical mapping of 3D computational topology to real hardware. Eliminating rackmount infrastructure also means potentially much higher processing density. Interconnections are not constrained by granularities relating to cards, racks, cabinets, and so-on.

TABLE I. SUMMARY OF IMPORTANT CHARACTERISTICS OF THE SURVEYED PACKAGING SYSTEMS

References	Topologies	Inter-board Packaging	Power delivery	Inter-board Communication	Hardware Implementation
[14]	optical multi-mesh hypercube	not specified	not specified	wireless (optical)	conceptual
[15]	hypercube and mesh	not specified	not specified	wireless (optical)	conceptual
HAEC [9]	wireless configuration	HAEC Box	not specified	wireless	HAEC playground (network-protocol evaluations)
[10,11,12]	wireless configuration	ball-shape object	wireless	wireless	conceptual
ExaNest [1,2]	hybrid [16]	rack/cabinet	backplane	wired	rack/cabinet
a variant in [13]	3D mesh, (4D hypercube at inter-processor level)	hexagonal-shape module, composed to a ball.	not specified	wireless	conceptual
this paper	3D mesh (3D torus with external data channels)	- as [13], investigating module's coupling and large-scale composition	3D power grid	direct via mated connectors	hexagonal board and frame prototype

III. RELATED WORK AND TCA DESIGN

Naturally, the tiled system has its own constraints, and its own unique properties. One of the most important is the notion of a decentralized power grid property, rather than parallelized backplane power bus, but there are others. Therefore, investigating the feasibility of such systems and understanding those properties and constraints is the key concern of this research. The goal is to determine if such systems are physically feasible when extended to large scale systems. Questions we particularly wish to answer in this research challenge include:

- Can a collective power grid sustain systems of large scale?
- Are we able to manage and predict power behaviors?
- Can such a system feasibly be physically constructed?
- Can workloads be varied node by node to optimize power distribution and computational throughput across a TCA?

Our work in some of these areas, as reported here, are a progression toward answering these questions individually and collectively.

In this section, we highlight relevant previous work (subsection A and B), and then detail the evaluated system design considerations.

A. Rack-mount Packaging

We briefly mention some parallel computers built with rack-based packaging in this category as it is considered a traditional inter-board level method. Recently, a number of projects have targeted large computing system challenges to achieve the next step of computing power at a minimum of billion-billion floating-point operations per second, i.e., exascale. ExaNeSt [1,2], ExaNoDe [3], ECOSCALE [4], and EuroEXA [5], are four example projects closely collaborating for the purpose. ExaNeSt focused on developing interconnection networks, storage, and cooling. The project employed the cooling system of ICEOTOPE [6]. The electronic circuit boards were submerged in warm non-conductive (dielectric) liquid flowing into and out of each of the blades contained in a rack. Another recent parallel computer was Supercomputer Fugaku [7]. The machine achieved the first rank in High Performance LINPACK (HPL) benchmark on TOP500 project [8], which was also built on rack-based packaging.

B. Non rack-mount Packaging

The packaging techniques in this subsection are more directly relevant to our work as they share some common configuration with our design. Thus, our work is considered a subset of this category. HAEC [9], was a project proposing a holistic energy-efficient computing system with both optical and wireless communication. In the project, a group of boards was named as HAEC Box. Another design of this category was conceptualized with a wireless computing system [10], to mitigate the complexity of data communication wiring, heat dissipation, power lines, and system composition effort. Afterwards, [11,12] further investigated the techniques. In [12], a level of abstraction of wireless interconnection network was designed for the concept. Dedicated simulation and visualization tools were also built to evaluate the performance of the wireless system behavior. It was concluded that at the time of the research, technologies of radio devices still consumed a large amount of energy, with improvements needed before this becomes viable. For performance analysis of [12], it was reported that a reasonable performance can be achieved on particular tasks executed on certain networks.

Subsequently, [13], proposed a variant of the concept. The packaging technique allowed cooling fluid to pass through a level of composition in order to dissipate heat from each unit. For the packaging in [13], we envisaged the feasibility of two alternative designs: both wired and wireless communication. For wireless communication, transceivers can be embedded in the smallest unit. On the other hand, in a wired design each edge of the unit can be used as an interfacing area for both data communication and power lines routed into the internal components.

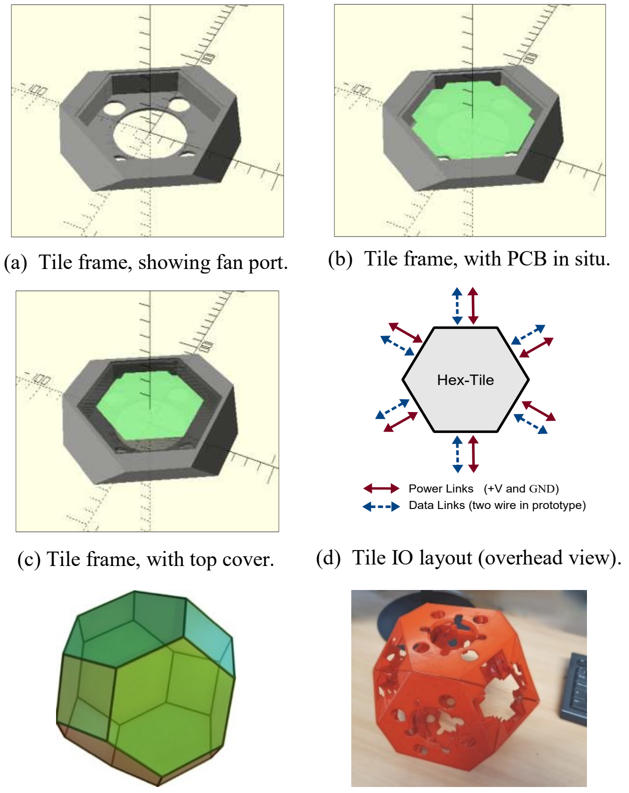
The power-route network enables a node to tolerate some faulty power-route situations. With a single unit added to the system, it provides the diversity of both powering and data communication networks. With such a method of powering nodes in the system, a challenge regarding electrical constraints emerges, which does not exist in traditional rack-based systems. A survey and comparison of related technologies is given in Table I. To investigate how practical the TCA is, in terms of physical scalability prior to a concrete implementation phase of a large system, work reported in this paper focuses upon wired communication for simplicity in our first investigation.

The TCA concept relies upon a fundamental building block – the 'hex-tile', and abstract views of which are shown in Fig. 1. Each tile is a hexagonal planar structure, with edges having alternate angles, as illustrated in Figs. 1a-1c. The space inside a tile may contain power-conversion units, computing, and communication elements such as CPU, memory unit, and a router, as illustrated by Fig. 1b. Power and ground lines and physical data channels can be routed via each of the six edges, creating the IO connectivity showing in Fig. 1d. IO lines typically act as independent point-to-point channels, while all tile power inputs are shared via common rails within each tile. Meanwhile, each tile is capable of joining to other tiles via the angled edge connectors, permitting a number of tiling schemes, including 2D planar tiling, and 3D topologies, including a ball-like structure comprising 8 tiles. Fig. 1e shows the shape when tiles are formed into a ball (a truncated octahedron, also known as a Kelvin Bubble [17] and Fig. 1f shows an actual equivalent prototype tile structure.

Balls are also permutohedra, thus forming tileable structures in 3 dimensions via the trapezoidal faces of the structure, as shown in Fig. 2d, and as shown in a cubic array of 3x3x3 balls as illustrated in Fig. 2a. It might be thought that simple cube-shaped modules are adequate. However, many 3D tiling topologies are possible only with a ball-like module, rather than a cube, and many of these facilitate continuously linked voids between balls, which is highly valuable for air or liquid coolant flow. For maximum densities, however, the space between tiled balls may be packed with a second 3D grid of balls, similarly inter-connected, as illustrated in Fig. 2b. This agrees with the principle of packed truncated octahedrons [17]. While single packed arrays have up to (n^3) nodes in cubic space, a doubly-packed array can approach almost twice that of a singular array for large dimensions of n, with up to $(n^3) + (n-1)^3$ nodes.

The outer balls of the array present trapezoidal connection points to be used in the most convenient power delivery arrangement. The most aggressive approach is to connect power and ground lines to all connectors available at the outer perimeter of the system. This allows the best-possible electrical current delivery and distribution throughout the system, but with a considerable degree of redundancy in power connections. The total number of connections could however be reduced significantly while maintaining a viable power grid.

Obviously, with the unique power network topology of a ball-grid, the voltage, current, and power delivery available to tiles in the different locations within a structure will vary to a degree, affected by connector-pin resistances, power consumption, and the overall collective power grid pathways. Moreover, connector-pin currents are also a special concern as the current-flow network are not obvious compared to those in rack-mount systems, and pin power/current carrying capacities have upper limits that must be respected. These concerns introduce a unique challenge for the electrical constraints, and thus ultimately a need to predict such behaviors within a dynamically work-loaded system.



(a) Tile frame, showing fan port.

(b) Tile frame, with PCB in situ.

(c) Tile frame, with top cover.

(d) Tile IO layout (overhead view).

(e) Truncated Octahedron. *https://en.wikipedia.org/wiki/Truncated_octahedron*

(f) Prototype tile frames assembled into Ball structure.

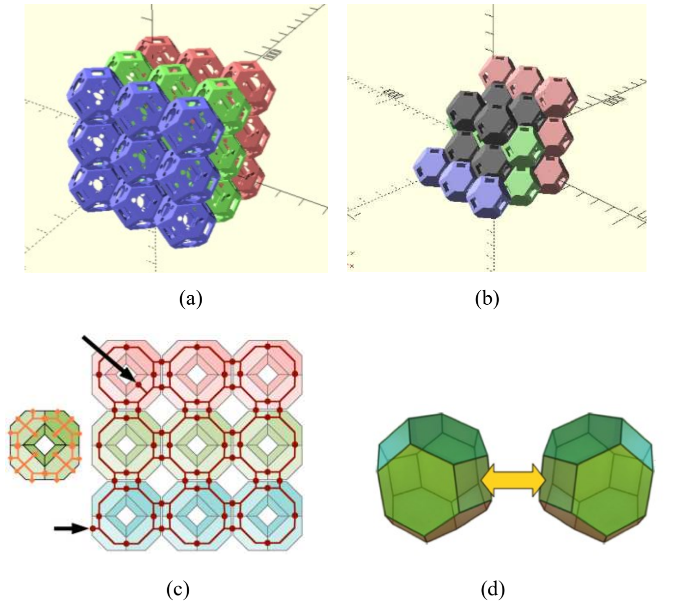
Fig. 1. Illustrations of the hex-tile. (a) shows a 3D model of the tile module frame, plastic or ceramic package material, (b) shows the tile frame with PCB or MCM in situ, (c) shows the prototype module top plate, (d) shows the IO connectivity of each tile edge, where solid/red arrows representing power and ground lines, and the dashed/blue lines are data channels, (e) shows the ball arrangement when 8 tiles are combined (forming a truncated octahedron with hexagonal and trapezoidal faces), and (f) shows an actual prototype tile-frame combination of 8 tiles into a ball (shows unpopulated tile frames).

C. Electrical Constraints

It is essential to ensure that all of the tiles in the system can operate without violating any electrical constraints, as defined within the specifications of their connectors and components. In this paper, we define three key constraints as follows.

- The tile onboard regulator output-load voltages are regulated at the specified levels.
- The tile input-voltages are in the operating ranges specified by the power-conversion units.
- The connector-pin currents do not exceed the levels specified by the manufacturer of the chosen connectors.

Practically, a system can be heterogeneously designed, composed of different tile types (SSD, Memory, CPU, FPGA, DSP, TPU to name a few). Thus, each tile may contain different load requirements, power-conversion units, and the limits of connector currents. In this paper, we assume all tiles are of uniform type, and the load resistances are steady, with constant power load and connector current limits at 3 A per single pin.


(a)

(b)

(c)

(d)

Fig. 2. (a) Visualized TCA system of 27 balls with the dimension of (3,3,3) in isometric view. (b) Example of ‘intra-grid packing’, 2 × 2 × 2 gray balls can pack in between the existing 3 × 3 × 3 grid (cutaway view). (c) Top view of array showing example power connection points (black arrows), notional power grid connectivity (right), and full connectivity paths (left). (d) Trapezoid edges as inter-ball connection points, or external power source interface points.

D. Models

Most power delivery models assume a bus and tree-like power distribution network, unlike the scheme employed in TCA. The nearest model in [18] is an example in our survey that holds a similar idea in terms of circuit components we expect for simulation. However, that power model is applied in the large-scale integration (VLSI) design level, where complex power grids are common. To give more details concerning our survey, in [19], a large amount of power and energy models related to HPC systems have been surveyed and classified in terms of system components. In that survey, we found that researchers paid more interest to the power modeling of either nodes, interconnects, or the whole system, rather than how power-delivery mediums are modeled. For this reason, we decided to design our own circuit model and simulation tool for our constraint evaluations, and ultimately to validate this against a real physical prototype. This is described as follows:

1) Pin-resistance model: Due to the cascading effect of connectors in the envisaged power grid, it is important to evaluate how the bulk conductor and contact resistances of connector pins impact on the scalability of the TCA system, thus a suitable model is required. In Fig. 3, the connectors, and their respective resistance models are depicted. Apart from the fairly constant bulk resistance of a pin, the contact resistance is also an important factor of the stability of the system, and can be measured or obtained from the connector datasheet. In our model, we use a single lumped resistor, named *r_p_resist* to model either a single tile-edge power pin, or collectively model multiple power-pins, if used in the same connector (power pins can optionally be doubled up in parallel to give higher current capacity).

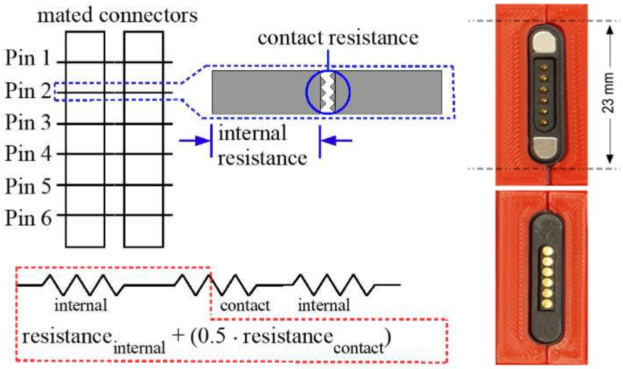


Fig. 3. Magnetic connector pair, and individual pin resistance modeling detail. Resistance *r_p_resist* (in the red, dashed frame) comprises the bulk internal pin resistance and a 50% share of pin-mating contact resistance as defined earlier.

Thus, a parallel-resistor calculation can be simply applied to assign a single resistance value to this *r_p_resist*. For a ground pin, an equivalent single resistor is named as *r_g_resist*. All of the currents passing through these resistors will be collated for connector-constraint evaluations. In this paper, all connector resistances are assumed to be the same. However, a more advanced approach could be considered in future, where variations could be modeled to represent localized pin resistance factors.

2) Board model: The inclusion of a switching regulator circuit model (LT3976 [20]) in our tile prototype results in excessively long simulation times for a large system. Thus, we sought a simplified model to evaluate the entire system in a steady state with constant regulator load(s). It was noted that [21] provides several average-model methodologies, and [22,23] also automate the modeling processes of an average model for switching regulators. It was determined that the curve fitting method was adequately effective for a simplified model to evaluate the system in its steady state while dramatically reducing simulation times (by a factor of hundreds), and file sizes, without significant loss of accuracy (typically less than 1% for tested cases). The simulator tools can select and use either approach according to accuracy and time constraints.

In Fig. 4, the abstract model, as a representation of the tile, can be depicted in the inner hexagon. the resistor at the center, *board_resistance*, represents the varying instantaneous equivalent-resistance of the entire board. The adjuster unit imitates the operation of a switching regulator, periodically samples both the input voltage, *vin_s*, and current, I_{board_s}, of the board. The adjuster adapts the value of *board_resistance* when a sampled board input-current is not “close enough” to the expected instantaneous input-current, I_{board_e}, as shown in (1). The parameter *Idiff_thres* (input-current difference threshold) controls this alignment, resulting in the accuracy of the simulation results. When the difference between the sampled and the expected input-currents, I_{diff}, shown in (2), is within the interval of ($-I_{diff\_thres}$, I_{diff_thres}), the adjuster maintains *board_resistance* value. Once every *board_resistance* in the system is stable, the entire system reaches the steady state. At this point of simulation, all the connector-pin currents, board input voltages, and currents can be collected for constraint evaluations. The parameter *tr_init* sets the period of the initial resistance before the step resistance, *Rstep*, takes the role of gradually altering *board_resistance*. For more rapid simulation, curve fitting may be used. Polynomial fitting of degree three was found to be adequate for our evaluations in this paper. The equations regarding the board model are given in (1) and (2). The equations of the board model can be implemented in an LTspice [24] simulation file using built-in symbolic sample-and-hold function blocks. In the simplified model, the initial resistance parameter may impact the time required for the LTspice simulator to achieve the DC operating point before every tile reaches the steady state. This can be seen at the simulation time of approximately 120 µs in Fig. 5.

$$\mathbf{I_{board\_e}} = p_1 vin\_s^3 + p_2 vin\_s^2 + p_3 vin\_s + p_4 \qquad (1)$$

$$\mathbf{I_{diff}} \quad = I_{board\_s} - I_{board\_e} \qquad (2)$$

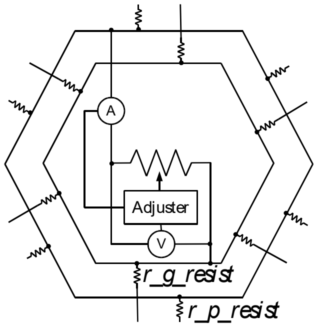
where:

vin_s	Sampled instantaneous board input-voltage
$P_{1..4}$	Coefficients of curve-fitting equation for a constant regulator-load power
I_{board_e}	Expected instantaneous input-current at steady state
I_{board_s}	Sampled instantaneous board input-current
I_{diff}	Difference between I_{board_s} and I_{board_e}

E. Model Validations

The system with the 'simplified' model was validated against the 'complex' LT3976 spice model, with a 3 × 3 × 3 ball array, and found to be averaging less than 1% margin of error for examined cases under load conditions. To simplify the validation model and shorten the simulation time, the soft-start mode of the tile power regulators was disabled, and load resistances were set to 1 Ohm, representing approximately 25 W per tile being regulated at 5V (25W being the maximum permitted for this particular regulator). A common 12 V external voltage source is supplied to all the system-surface power and ground pin models. The initial value of the external (surface) was set at 0 V, then ramped up to 12 V. This allows the LTspice simulator to more quickly achieve a DC operating point. Example parameter values and the LTspice code, as a part of adjuster unit are presented in Table III.

As noted in Table II (a) and (b), the complex model is compared to real prototypes for a single tile and an eight-tile ball. Simulator and hardware prototype results were found to closely agree in these, with typical agreement within the region of 1-2% for all simulated power-load cases (i.e., excluding no-load). Voltage stability across sample tile networks, as given in Table II (c), was excellent, and well below 0.5% where tiles are composed as a 2D or 3D tiled cases tested. As expected, group-tiled arrangements are more stable due to the parallelism and sharing of current paths across the power grid.



Dynamic power load options.

Mode (bin)	**Load R** Ohms	**Power** Watts
000	No load	0
001	10.00	2.5
010	5.00	5.0
011	3.33	7.5
100	2.50	10.0
101	2.00	12.5
110	1.67	15.0
111	1.43	17.5

Fig. 4. Conceptual representation of tile model. As per the real tile, 'power consumption' above base load can be dynamically adjusted via a CPU-selectable load resistance. Additional CPU load (regulator 5V output ~ 25mW). Tile cooling fan (~60mA, 12V rail, ~700mW) is separately modeled.

TABLE II. ACCURACY OF THE 'COMPLEX' LT3976 MODEL SIMULATION, VERSUS ACTUAL PROTOTYPE AND SELECTED SYSTEM CHARACTERISTICS.

(a) Prototype/Model: Single tile, Single connector

	Min (base)	**Low**	**Med**	**High**	**Max**
	~ 0W	+2.5W	+5.0W	+10.0W	+17.5W
$\boldsymbol{I_P \pm 5mA}$	60 mA	310 mA	540 mA	1000 mA	1760 mA
$\boldsymbol{I_M}$	62.29 mA	310.82 mA	539.93 mA	1012.81 mA	1753.82 mA
Error *(ave)*	4.5%	0.3%	0.0%	1.3%	-0.4%
(min, max)	13.3%,-4.2%	1.9%, -1.3%	0.9%, -0.9%	1.8%, 0.8%	-0.1%,-0.6%

(b) Prototype/Model: 8-tile ball, 2 co-located power connectors

	Min (base)	**Low**	**Med**	**High**	**Max**
	~ 0W	+2.5W	+5.0W	+10.0W	+17.5W
$\boldsymbol{I_P \pm 5mA}$	530 mA	2550 mA	4370 mA	8070 mA	14010 mA
$\boldsymbol{I_M}$	501.67 mA	2493.57 mA	4328.48 mA	8121.95 mA	14079.9 mA
Error *(ave)*	-5.3%	-2.2%	-0.9%	0.6%	0.5%
(min, max)	-4.4%, -6.2%	-2.0%, -2.4%	-0.8%, -1.1%	0.7%, 0.6%	0.5%, 0.5%

(c) Prototype: grid stability (worst case voltage drop, 10W load, 12V supply)

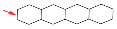
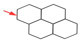
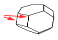
Tiling	**Configuration**	**Prototype**
	**1D tiling**: *4 tiles, 1 connector*	1.25%, 150mV
	**2D tiling:** *4 tiles, 1 connector*	0.33%, 40mV
	**3D tiling:** *8 tiles, 2 connectors*	0.17%, 20mV

TABLE III. LTSPICE EXAMPLE PARAMETERS

Example parameter values:

Initial resistance period: 6 Ohms, held for 21 us, then 0.005 Ohm steps

Example LTspice code with the above parameter values

```
b_i_board i_board v = i(r_board_resistance)
b_i_diff i_diff 0 v = v (i_board_s) - ( (-0.006025)*(v(vin_s)**3) +
   +  0.2087*(v(vin_s)**2) - 2.623*v(vin_s) + 14.39 )
b_r_board r_board 0 v = if(time<21us, 6 ,if( v(i_diff) > 0.01, v(r_board_s)
   + 0.005,  if(v(i_diff) <-0.01, v(r_board_s)-0.005, v(r_board_s))))
```

As shown in Fig. 5a, after the external voltage source reaches 12 V, the board input-voltages are at certain voltages. All the voltages are below 12 V, affected by the resistances of connectors located in different layers of the system. At this point, both the detailed and simplified models, Fig. 5a and 5b respectively, continue to converge into the steady state, with both models very similar at 120-140 µs.

F. Simulation Framework

In this paper, we focus on the feasibility of TCA, however, we also briefly describe the simulation framework to demonstrate how instances of the tile model can be composed into a complete system. To evaluate a large system means that a hierarchically complex resistor-network model needs to be generated and manually creating an LTspice simulation file is a tremendously labor-intensive task. Thus, we automate this process by building our own source-code file generators, which can generate a complete simulation model for any set of ball array dimensions. The automation of LTspice code generation starts at the inter-ball level of composition, generating a structure according to the required topology (in this case a cubic array of the type illustrated in Fig. 2a).

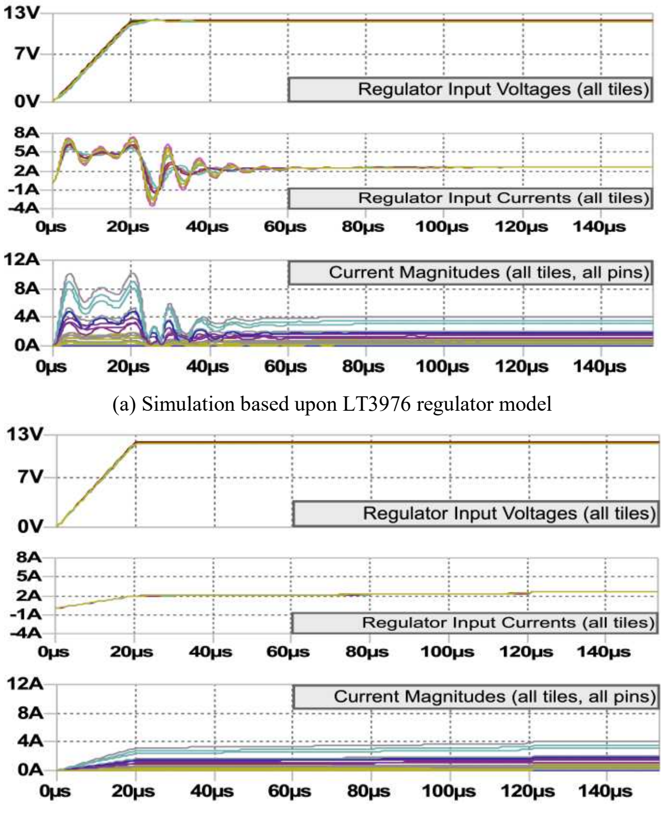
IV. Simulation Results

The number of balls in each dimension is parameterizable in our simulation framework, thus, arbitrary ball sizes of system can be generated. However, in this paper, only cube-shaped systems with 50 mOhms mated pin-pair resistance, with a single 12 V power source common to every surface connector, is evaluated and reported. Given that the individual pin current-limit is 3 A, power and ground pins are configured as doubled-up pairs, to permit up to 6 A. Fig. 6a and 6b show simulation results for multiple ball-array configurations, ranging from a single ball up to an $(n \times n \times n)$ array size of $n = 5$, with 125 balls and 1000 tiles. Power loadings per tile are varied from 5W to 25W (on the regulator output side).

It can be observed in Fig. 6a that, as expected, voltage drops across the power grid of each array will scale up as the cascaded effects of tile-to-tile pin connection resistances accumulate. In Fig. 6b, the maximum observed pin-currents across the grid are presented for the same range of ball-array configurations. Here it is observed that pin currents remain within the specifications of $I \leq 6$ A, until the cubic ball dimension reaches $n = 5$, at which point the pin current is exceeded in one or more pins across the array (for all tiles at full power load). However, this may well be happening in only a limited number of pins, and by moderating the power consumption on a tile-by-tile basis, for instance where some tiles operate at perhaps 20W rather than 25W, it should be possible to return pin currents to within specified limits.

An important advantage, therefore, of the availability of a modeling and simulation framework, is that it permits more advanced power management strategies to be explored. For example, a predictive power optimization model and visualization tool, based upon a genetic-algorithm (GA), has already been implemented. This GA optimizes the power allocated for consumption by each tile whilst achieving a target system-wide power load and ensuring connector current constraints for the whole system are within specifications.

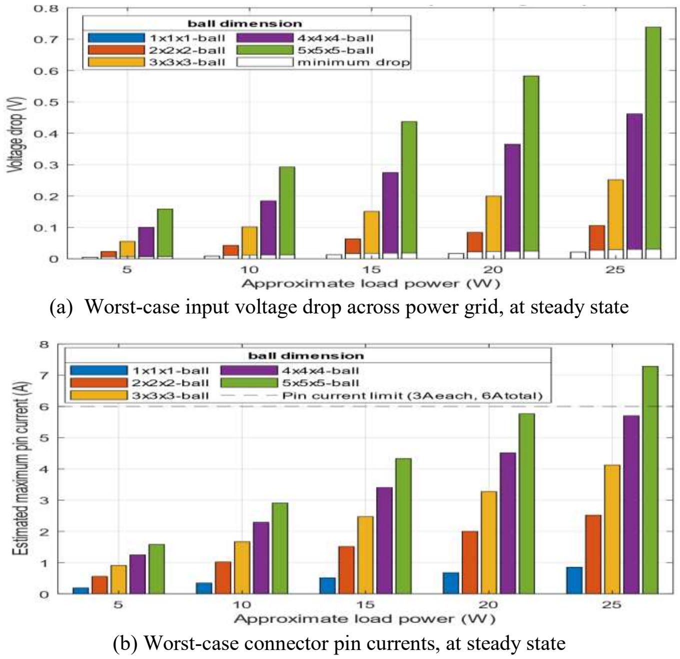
In effect, optimizing power per tile, loosely equates to optimizing computational capacity within the system for a given set of constraints, therefore, we may optimize array performance in this way.



(a) Simulation based upon LT3976 regulator model

(b) Simulation using simplified (faster) model

Fig. 5. Validation of the simplified board steady-load model with 3x3x3-ball system (27 balls, 216 tiles), showing (a) the LT3976 model, and (b) with the simplified (much faster) simulation model.



(a) Worst-case input voltage drop across power grid, at steady state

(b) Worst-case connector pin currents, at steady state

Fig. 6. Constraints simulation results (with 101mA assumed supply side 12V fan load in this case). (a) Estimated maximum board input-voltage drop and (b) Estimated maximum pin-currents for different load-power allocations and system sizes.

For example, Fig. 7 shows the visualization of the result of the experimental GA power optimization. It is observed that initial power loading distribution, as shown in Fig. 7a has been significantly improved in GA-enhanced loading of Fig. 7b, after GA algorithm converged to within a set margin of optimality.

V. Prototype System

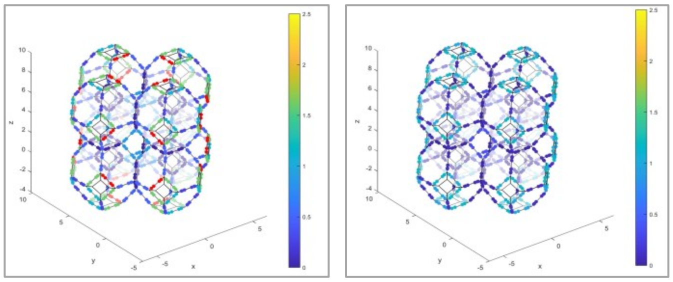
To explore and validate feasibility, and simulator accuracy, a hardware prototype has been developed, examples of which are showing in Fig. 8 in various levels of assembly and operation. Each prototype tile utilizes an LT3976 power regulator, onboard ATMEGA324PB microcontroller, acting mainly as a 'house-keeping' control node, data IO intermediary, and also able to dynamically control a dummy power load, emulating heavier power usage at the tile level. Magnetically coupled 6-pin power/IO connectors (as shown in earlier Fig. 3) permit tile-to-tile connection, with two IO lines, two positive supply pins and two ground rails (to achieve 6A current capacity). Current prototypes include a complete 8-tile ball, a base mounting platform, and relocatable surface power leads.

The system has been tested with dynamic power ranging across tiles up to maximum system power loading. Fig. 8d shows a snapshot (from video) of a ball (two tiles removed to permit interior view) under test conditions with power loading dynamically stressed across the grid. Onboard cooling for these prototypes is achieved via an air-flow fan (visible in Fig. 8a). At Maximum power load (17.5W for prototype power configuration options), with 14 A supplied to the ball, an interior air-space temperature of around 15c above ambient (~21c) was observed. This power loading could also be achieved by hosting a suitable CPU in the extension socket, with similar results.

VI. Further Hardware Development Directions

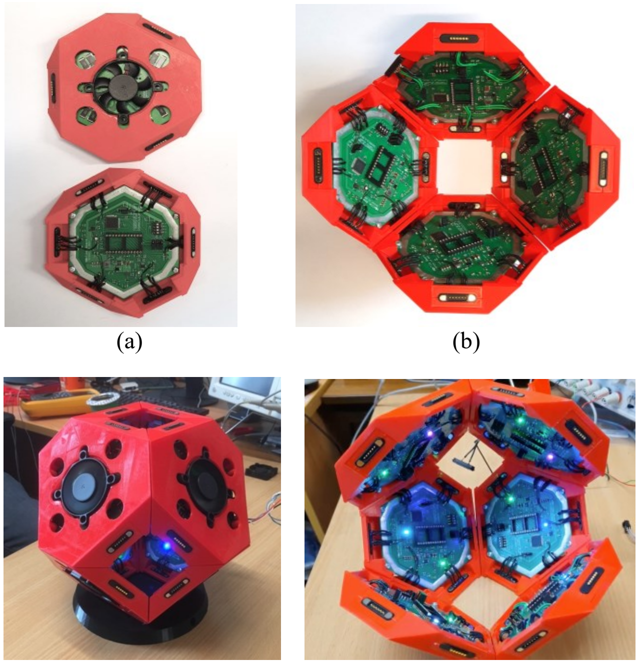
The prototype is necessarily over-sized given the construction methods available (individual hex-tiles are 120mm edge to edge, and a ball is approximately 200mm face to face). Dimensions of the order of 50-60 mm per ball are feasible when utilizing ceramic chip packaging technology to encapsulate single SOC or MCM modules representing processors, SSD, memory. A further option would use selected balls as power reservoirs to improve power availability under highly transient conditions and/or where local power demands within the grid change dynamically.

There are also possibilities to manufacture the balls as complete components and use these as the fundamental building blocks, with the same principles applying at a coarser granularity. Combination with liquid cooling systems would then be envisaged, as investigated in previous related work [6,10,11]. At this level of physical size, individual tile cooling would be dropped, and air/fluid flow-assistance via inter-ball modules located and interspersed at the trapezoid connectors would permit controllable dynamic (fluid or air) flow control across any array topology constructed. This concept is illustrated in Fig. 9. Notably, even in the case of the double-packed array of Fig. 2b, the cooling model still supports appropriate capacity to remove heat, since the cooling network is duplicated as two independent flow networks in the two interleaved arrays, with proportionate increase in cooling.



(a) Initial conditions (b) Improved loading

Fig. 7. 3D edge-connector current visualization for a 2x2x2 ball array (64 tiles). The colored dots represent edge-connector pin currents (colorized blue through to yellow for normal loadings). Red dots highlight exceeded pin current locations. The genetic algorithm achieves better power distribution within the grid by changing the power utilization on each tile while maintaining the overall target power consumption (and thus computational capacity).



(a) (b)

(c) (d)

Fig. 8. (a) Hex-tile prototype (top and reverse views), (b) Four hex-tiles linked into a half-ball (petal) formation, (c) Eight tile-frames comprising a ball with a base-plate, with trapezoidal connection faces visible, (d) A ball, powered-up with shared power distribution between tiles (top two tiles removed for interior view, LED colors relate to power loading).

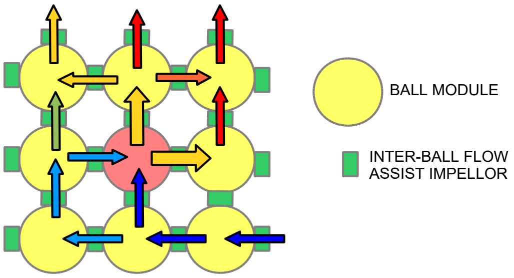


Fig. 9. Inter-ball fan/pump/impellor and air/liquid flow control principle. This shows a 2D view, but in practice would operate in 3 dimensions to modulate thermal flow dynamically under system monitoring and control. Cool air/fluid flows into the grid (blue) accumulates heat transfer in a directed fashion according to localized need, and exits system (Red). Existing prototypes can already operate in a similar mode, though in a less optimal fashion.

VII. Conclusion and Future Opportunities

An investigation of a novel concept for extensible 3D processor array topologies has been presented, comprising of a novel hexagonal tile design, permitting the assembly into modular truncated octahedron 'ball' modules, and which may be combined into larger scale arrays in a variety of topologies. It has been demonstrated that power delivery within the unusual structure, and particularly the distributed power mesh, is a workable model and may be effectively predicted and managed.

A physical prototype has been briefly described and demonstrates that the system concept is practically realizable. The behavior of the prototype hardware was found to be typically within a few percent of simulation predictions, suggesting that the simulation model is representative of similar systems at larger scales, and that engineering constraints involving power and current densities may be identified and managed appropriately.

This work, alongside others [10,11,12,13] takes an important step toward the realization of large-scale systems based upon tiled modules without traditional rack-mount architecture overheads and constraints. To progress further there are several avenues for this concept to be pursued. The use of optimal workload balancing across a topological array, in order to manage optimal power distribution versus workload throughput, dynamic power and thermal management strategies, and the exploration of thermal management technologies including airflow and fluid systems. Communications channels are currently physical point to point. However, work has already been done in the field in relation to short near-field communications at high data rates using localized wireless data links, with point-to-point, multicast and broadcast potentials.

Meanwhile, the level of integration and physical size of the hex-tile requires a further step-change. Ultimately, the basic building block may be a smaller tile, or a complete ball on smaller scales. Such modules would likely utilize relatively well-established manufacturing technologies: Ceramic chip packaging materials and custom chip-carrier designs, employing single-chip systems with complete processors, memory, storage, routing.

As these areas are advanced incrementally, the authors expect to see feasibility of large-scale tiled arrays becoming greatly improved, ultimately moving toward realizable commercial systems.

Acknowledgements

1. This work was supported by Royal Thai Government NSTDA PhD scholarship (OEA ID No. ST_G5599).
2. The authors also wish to thank the University of York, Department of Computer Science, technical team for their assistance with 3D printing and board assembly work.

References

[1] M. Katevenis et al., "The ExaNeSt Project: Interconnects, Storage, and Packaging for Exascale Systems," 2016 Euromicro Conference on Digital System Design (DSD), 2016, pp. 60-67, doi: 10.1109/DSD.2016.106.

[2] R. Ammendola et al., "The Next Generation of Exascale-Class Systems: The ExaNeSt Project," 2017 Euromicro Conference on Digital System Design (DSD), 2017, pp. 510-515, doi: 10.1109/DSD.2017.20.

[3] A. Rigo et al., "Paving the Way Towards a Highly Energy-Efficient and Highly Integrated Compute Node for the Exascale Revolution: The ExaNoDe Approach," 2017 Euromicro Conference on Digital System Design (DSD), 2017, pp. 486-493, doi: 10.1109/DSD.2017.37.

[4] I. Mavroidis et al., "ECOSCALE: Reconfigurable computing and runtime system for future exascale systems," 2016 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2016, pp. 696-701.

[5] "EUROEXA." euroexa.eu. https://euroexa.eu (Accessed Jun. 30, 2022).

[6] "Precision immersion cooling from the Cloud to the Edge | Iceotope." iceotope.com. https://www.iceotope.com (Accessed Jun. 30, 2022).

[7] "Fugaku | RIKEN Center for Computational Science RIKEN Website." https://www.r-ccs.riken.jp/en/fugaku (Accessed Jun. 30, 2022).

[8] "November 2021 | TOP500." top500.org. https://www.top500.org/lists/top500/2021/11 (Accessed Jun. 30, 2022).

[9] G. P. Fettweis et al., "Architecture and Advanced Electronics Pathways Toward Highly Adaptive Energy- Efficient Computing," in Proc. of the IEEE, vol. 107, no. 1, pp. 204-231, Jan. 2019, doi: 10.1109/JPROC.2018.2874895.

[10] J. Austin (Cybula Ltd), "Computing Devices" GB Patent No. GB02/04104, September 10, 2002.

[11] R. Hind, "Feasibility Study on implementing the "Ball Computer"," M.S. thesis, Dept. Comput. Sci., Univ. of York, York, 2013.

[12] A. M. Kamali Sarvestani, "Evaluating Techniques for Wireless Interconnected 3D Processor Arrays," Ph.D. thesis, Dept. Comput. Sci., Univ. of York, York, 2013.

[13] A. M. Kamali Sarvestani, C. Crispin-Bailey, and J. Austin, "Performance Analysis of a 3D Wireless Massively Parallel Computer," J. Sensor and Actuator Networks, vol. 7, no. 2, pp. 1-20, Apr. 2018, doi: 10.3390/jsan7020018.

[14] A. Louri and H. Sung, "An optical multi-mesh hypercube: a scalable optical interconnection network for massively parallel computing," in Journal of Lightwave Technology, vol. 12, no. 4, pp. 704-716, April 1994, doi: 10.1109/50.285368.

[15] A. Louri and Hongki Sung, "3D optical interconnects for high-speed interchip and interboard communications," in Computer, vol. 27, no. 10, pp. 27-37, Oct. 1994, doi: 10.1109/2.318581.

[16] J. Navaridas, J. Lant, J. A. Pascual, M. Luján, and J. Goodacre, "Design Exploration of Multi-Tier Interconnection Networks for Exascale Systems," 2019. doi: 10.1145/3337821.3337903.

[17] W. Thomson. "On the division of space with minimum partitional area," Acta Mathematica, vol. 11, pp. 121–134, Mar. 1887.

[18] Z. Zhang, X. Hu, C. Cheng and N. Wong, "A block-diagonal structured model reduction scheme for power grid networks," 2011 Design, Automation & Test in Europe, 2011, pp. 1-6, doi: 10.1109/DATE.2011.5763014.

[19] K. O'brien, et al, "A Survey of Power and Energy Predictive Models in HPC Systems and Applications," ACM Comput. Surv., vol. 50, no. 3, Jun. 2017, doi: 10.1145/3078811.

[20] Linear Technol. Corporation, Milpitas, CA, USA. *LT3976 - 40V, 5A, 2MHz Step-Down Switching Regulator with 3.3µA Quiescent Current*, (2013). Accessed: Jun. 30, 2022. [Online]. Available: https://www.analog.com/media/en/technical-documentation/data-sheets/3976f.pdf

[21] C. P. Basso, Switch-Mode Power Supplies, 2nd ed. New York, NY, USA: McGraw Hill Education, 2014.

[22] M. H. Leonard, "Automated Behavioral Modeling of Switching Voltage Regulators," B.S. Thesis, Dept. Elect. Eng., Univ. Arkansas, USA, 2013.

[23] M. H. Leonard, "Semi-Automated Switching Regulator Modeling Method and Tool," M.S. Thesis, Dept. Elect. Eng., Univ. Arkansas, Fayetteville, AR, USA, 2015.

[24] "LTspice Simulator | Analog Devices." analog.com. https://www.analog.com/en/design-center/design-tools-and-calculators/ltspice-simulator.html (Accessed Jun. 30, 2022).