Communication Architecture Enabling 100x Accelerated Simulation of Biological Neural Networks

Kevin Kauth, Tim Stadtmann, Ruben Brandhofer, Vida Sobhani and Tobias Gemmeke
IDS, RWTH Aachen, Germany
kauth@ids.rwth-aachen.de

ABSTRACT

To further develop the understanding of cognitive processes in the human cortex, neuroscientists seek to simulate relevant biological neural networks in the order of 10^9 neurons with natural densities of 10^4 synapses per neuron. To observe long-term effects of learning, a speed-up of at least 100x with respect to biological real-time is required while preserving deterministic results and a high temporal resolution of 0.1 ms. In this paper, we translate these objectives to requirements for the communication architecture of a large-scale neuroscience simulator. These requirements are based on a connectivity model that includes gray and white matter as well as clustered connections and represents essential communication requirements of biological neural networks. In analytical and numerical analysis, existing platforms fall short of meeting all requirements simultaneously even assuming modern high-speed transceivers. This paper presents a balanced multi-hop communication architecture that cuts latency and achieves high bandwidth efficiency. Extrapolating from physical measurements of link performance, our work brings the challenging communication requirements within reach of next generation large-scale neuroscience simulation platforms.

CCS CONCEPTS

• **Networks** → **Network design principles**; • **Hardware** → *Buses and high-speed links*; • **Computing methodologies** → *Parallel computing methodologies*; • **Computer systems organization** → *Grid computing*; **Special purpose systems**.

KEYWORDS

Neuromorphic Computing, Biological Neural Networks, Spiking Neural Networks, Grid Computing, Communication Architecture, Network Topologies

ACM Reference Format:
Kevin Kauth, Tim Stadtmann, Ruben Brandhofer, Vida Sobhani and Tobias Gemmeke. 2020. Communication Architecture Enabling 100x Accelerated Simulation of Biological Neural Networks . In *System-Level Interconnect - Problems and Pathfinding Workshop (SLIP '20), November 5, 2020, San Diego, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3414622.3431909

SLIP '20, November 5, 2020, San Diego, CA, USA

ACM ISBN 978-1-4503-8106-2/20/11.
https://doi.org/10.1145/3414622.3431909

1 INTRODUCTION

The cognitive capabilities of the brain stimulate research across many scientific fields. With the increase in computing performance, artificial neural networks have surpassed human level capabilities in specific tasks. However, the cognitive computing principles of the neocortex still remain a mystery. After the Human Brain Project triggered a first generation of large-scale neuroscience simulators [16], it appears that a next system generation is required to dissect the intricacies of dynamics on the neuron level and at large scale.

Neuroscientists envision to simulate learning processes of $N = 10^9$ neurons in a connectome, accelerated by a factor of $a = 100$x [13]. Within a biological time step of $h = 0.1$ ms motivated by minimal axon delays, the accelerated simulation needs to finish computation and communication within 1 µs. Scientific experiments therein require *deterministic* reproduction, *flexibility* in modelling and detailed *observability* down to the level of individual membrane potentials. No state-of-the-art system fulfills all of these at once.

Out of the many challenges, this paper focuses on the conception of a communication architecture that meets the implied requirements in terms of bandwidth and latency. As a baseline, we consider a distributed system consisting of N_N interconnected compute nodes, each simulating NpN neurons of a squared, disjoint section of neuronal tissue. We start by defining a biologically-plausible connectivity model including white and clustered gray matter connections in Section 2. State-of-the-art communication concepts and systems are discussed in Section 3 and 4, respectively. In Section 5, we detail quantitative assessments using analytical and numerical techniques to evaluate the feasibility of said architectures. Finally, we introduce a novel multi-hop architecture and multi-stage routing scheme to close the gap in target performance, as presented in Section 6 and 7. We draw our conclusion in Section 8.

2 BIOLOGICAL SPECIFICATION

To evaluate system performance of communication architectures for neuromorphic simulation, we need to create a representative benchmark. Therefore, we begin our analysis by developing a model of the neuronal connectome based on biologically realistic, conservative assumptions which place plausible demands on communication (cf. Tab. 1 for a summary of the resulting model).

For analyzing communication requirements, a neuron can be treated as a black box absorbing incoming and generating outgoing spikes with a specific rate and target distribution. The time of generation and the presynaptic neuron ID represent all relevant spike information. This allows action potentials to be expressed in an address event representation (AER). However, some communication schemes might require additional information as reasoned below.

The network load is proportional to the rate of signaling events. Various studies have shown that *firing rates* of cortical neurons

span multiple orders of magnitude with a log-normal distribution, ranging from less than 0.01 Hz to over 100 Hz, depending on neuron type, recording strategy, etc. [3]. Both *in-vivo* and *in-vitro* experiments suggest this distribution to be skewed towards its lower end - the majority of cortical neurons seem to be "silent", while a minority is responsible for most firing events [3, 26]. Regarding a network of 1 billion neurons and the accompanying immense number of firing events, the law of large numbers suggests the average firing rate to converge to its expected value. This value is bound by the metabolic cost of firing to around 7 ± 2 Hz [29]. We therefore assume a cortical *mean* firing rate per neuron of $\nu = 10$ Hz as a conservative estimate for upper-bound analyses of the communication load later on. This is in line with previous studies [1, 7, 23].

Asides the firing rate, the number of *synapses per neuron* can have substantial impact on network load. It defines the fan-in and -out of each neuron, and therefore the number of targets each generated spike has to be transmitted to. While the exact count per neuron can vary, the average number of synapses per cortical column appears to be $10^3..10^4$ [5, 20]. Therefore, we adopt the conservative estimate of $SpN = 10^4$, following previous studies such as [7].

Beyond throughput, communication latency is the critical limitation for achieving high acceleration factors. The biological *propagation speed* v over an axon varies: for unmyelinated axons (gray matter) $v_\mathrm{g} \leq 2.2\,\mathrm{m/s}$ [30, 31], and for myelinated axons (white matter) $v_\mathrm{w} \leq 60\,\mathrm{m/s}$ [14, 29, 30]. The latter is bound by a minimal latency of 1 ms [29]. Together with the number of *neurons per node* NpN and the *neuron density* in the neocortex $\rho_\mathrm{N} \approx 77 \times 10^9/\mathrm{m}^2$ [21], these define the maximal system latency in terms of hops per simulation step. NpN is a design parameter that depends on the capabilities of a node and the model complexity. In our analyses, we assume that neurons are mapped to compute nodes in a way that preserves their spatial distribution, following their biological density. First assertions on off-the-shelf FPGAs resulted in a number of neurons per node in the lower thousands. When not mentioned otherwise, we therefore set $NpN = 10^3$. A sensitivity analysis regarding NPN will be conducted in Section 7.

Finally, we regard the *spatial distribution of postsynaptic targets* for a single neuron. A basic, worst-case example would be a uniform target distribution across all neurons. As closer neurons have higher probabilities to form synapses, slightly more realistic, yet still simplified models have been introduced as well, e.g. taking into account the *locality* of spikes [18]. Considering actual biological distributions, these models neglect two important aspects of neuronal connectivity: 1) neurons tend to have clustered groups of targets, and 2) white matter connections enable high-speed communication between neurons. We expect both to have a large impact on bandwidth and latency constraints which has to be addressed and possibly leveraged by suitable communication architectures.

Based on these insights, we developed the target distribution model sketched in Fig. 1. It is an extension of the model of [28] that additionally accounts for the demanding high speed white matter connections. All in all, it integrates the following three distributions:

(1) central gray matter cluster with 40% of all targets (2D normal distribution with SD $\sigma = 0.297$ mm, bounded by r_c) [24, 28]
(2) 3 non-central gray matter clusters with 20% of all targets, centers between r_1 and r_2, and radii of $r_\mathrm{p} < r_\mathrm{c}$ (due to lack of details in literature we presume the worst-case: a uniform distribution within these clusters) [24, 28]
(3) long-range white matter connections to 40% of all targets, starting at r_w (exponential distribution with $\lambda = 0.11/\mathrm{mm}$, for simplicity without cutoff) [24, 27]

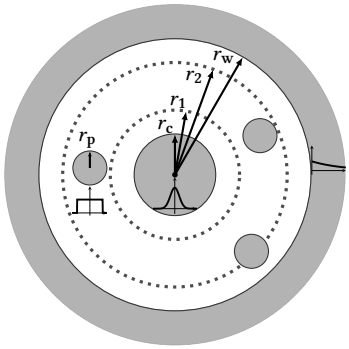

Figure 1: Connectivity model ($r_\mathrm{c} = 0.5$ mm, $r_1 = 0.75$ mm, $r_2 = 7.75$ mm, $r_\mathrm{w} = 8$ mm, $r_\mathrm{p} = 0.25$ mm).

The parameters derived in this section constitute an abstract model of neuronal connectivity in the human neocortex, suitable to pinpoint hardware bottlenecks. The heavy-tailed distributions of connectivity and firing rates, additional modes of information flow like neuromodulation, gap junctions and glial cells, and the isotropic view at neurons across layers are in their entirety not negligible for understanding cognitive processes in the human brain. However, as we aim to model communication requirements, spikes have the most substantial impact on quantities like bandwidth and latency. Transitioning from these biologically-reasonable estimations to more realistic models can be assumed to not change the main deductions of this work, as their effect on communication is either minor or only relevant on computations. The complexity of computation directly translates into the number of neurons per node, whose effect on communication will be examined later.

Table 1: Considered system requirements

Description	Variable	Value	Unit
Biology			
Total number of neurons	N	10^9	
Synapses per neuron	SpN	10^4	
Neuron density	ρ_N	$77 \cdot 10^9$	$1/\mathrm{m}^2$
Firing rate	ν	10	Hz
Max. velocity in gray matter	v_g	2.2	mm/ms
Max. velocity in white matter	v_w	60	mm/ms
Simulation			
Simulation step biological time	h	0.1	ms
Acceleration factor	a	100	
Implementation			
Message size	l_m	128	bit
Neurons per node	NpN	1000	
Number of nodes	N_N	N/NpN	

3 BACKGROUND

After discussing the biological parameters that define the base requirements for the targeted system, we now turn to existing concepts for communication architectures and technical capabilities of modern transceivers.

3.1 Technical capabilities

Error rate. State-of-the-art wireline communication typically targets a bit error rate (BER) of $p_\mathrm{b} \leq 10^{-15}$. Applying this to a neuromorphic simulator using a broadcasting scheme, for instance, spike packets of $l_\mathrm{m} = 128$ bit lead to $N \cdot N_\mathrm{N} \cdot a \cdot \nu \cdot l_\mathrm{m} \cdot p_\mathrm{b} = 128 \cdot 10^3$ packet errors per second. Screening packets with n-bit cyclic redundancy check (CRC) reduces the mean time between failures (MTBF), which is desirable in a reproducible simulation platform. Assuming n=32 bit, which leaves 2^{-n} undetected errors [22], the MTBF reduces to 10 hours. Including packet header and payload, this confirms the aforementioned $l_\mathrm{m} = 128$ bit.

Bandwidth. Modern high-speed NRZ transceivers reach up to 32 Gbit/s (e. g. [6]). In modern FPGAs, the resulting aggregated bandwidth exceeds 8 Tbit/s (e. g. [32]). Amongst others, this is achieved by employing more complex modulation schemes that reach higher throughput while increasing latency. On top of that, standardized communication protocols increase network load due to protocol overhead (e. g. Ethernet +400% for 128 bit packets).

Latency. Ethernet and Infiniband standards move towards a latency of 500 ns [17]. However, electrical signal propagation of local system interconnects is below 1 ns. As a link bandwidth of 32 Gbit/s only adds latency of 8 ns per packet, the rest can be attributed to on-chip processing, for example due to protocol overhead. For instance, hop latency between FPGAs can be reduced from 400 ns to 200 ns switching from Ethernet to a light-weight custom protocol [15]. We assert that current technical capabilities pose a lower latency bound of around 100 ns, considering a high-speed ASIC design in the unloaded case - the impact of network loading is covered in our dynamic simulations, as will be detailed in Section 7.

We use the *NetFPGA SUME* board [33] as a representative vehicle to validate the assessed technical capabilities. It features 32 high-speed transceivers supporting a bandwidth of up to 13.1 Gbit/s each, leading to an aggregated peak bandwidth of almost 420 Gbit/s. We conducted latency, bandwidth and error rate measurements using the Xilinx *Aurora* core with different encodings, frequencies and SATA/SFP+ cables of varying length ($\leq$ 1.5 m) and brand. In 8b10b encoding, a latency of 140 ns is achieved at a data rate of 6 Gbit/s. Larger packets using 64b66b encoding almost quadruple the latency to 540 ns, while higher data rates as well as on-chip frequency scale latency proportionally down to 270 ns at 12.5 Gbit/s. Finally, long-running error rate measurements using SATA3 cables at 6 Gbit/s and SFP+ cables at 12.5 Gbit/s confirmed a BER of 10^{-15}.

3.2 Communication architectures

To structure the following evaluation, we introduce three dimensions that span our design space of communication architectures.

Network topology. A network is composed of nodes and edges. Here, each node contains all resources necessary to compute the dynamics of NpN neurons and has a number of edges to other nodes, commonly referred to as (out-)degree. The communication latency in such a system is measured in hops, and therefore proportional to the time to cross a direct physical link between two nodes.

Common network topologies for high-performance computing are meshes (degree 4, 6 or 8), trees (binary, fat), and hypercubes. We disregard the alternatives provided by graph theory because of their complex implementation. To further reduce the exploration space, we take a closer look at the established topologies in terms of traffic balance and resource impact. Firstly, binary trees fail our litmus test regarding bandwidth and latency as they generate imbalanced traffic and excessive worst-case latency of biologically neighbouring nodes. Fat-trees aim at balancing traffic in trees at the expense of numerous resources without improving on latency. A corresponding problem exists for hypercubes with their inhomogeneous structure. Finally, 2D meshes create bandwidth bottlenecks towards their center. However, this is fully addressed with toroidal connections. Meshes can be improved further by additional link insertions [19]. We therefore limit the exploration to toroidal meshes of varying degrees D, referred to as *MeshD* in the following.

Casting scheme. Different messaging schemes can fan-out the action potentials from presynaptic (*source*) to postsynaptic neurons (*target*s). Unicasting (UC) sends SpN messages, one per spike and target. In contrast, broadcasting (BC) sends a single message per spike to all nodes. Ideally, multicasting (MC) sends single messages along shared paths with on-the-fly duplication in points of divergence, striking a source-side balance between UC and BC.

Routing algorithm. Routing paths can either be pre-computed (*offline*) and contained in message headers [1] or local routing tables [7], or resolved *on-the-fly* in the router. Efficient heuristics in the latter case are dimension order (DO) and longest dimension first (LDM) routing. The former is a deadlock-free, turn-restricted algorithm with a predefined direction order. [9]. The latter prioritizes the longest dimension [4]. In all examples, we assume the system to employ load balancing and properly sized buffers for deadlock prevention. As reference, we also consider the Dijkstra (DJ) algorithm which greedily pre-computes a shortest possible path.

4 STATE-OF-THE-ART SYSTEMS

There is a large body of literature in the field of neuromorphic computing [25]. Neuroscience software simulators like NEST:: [8] operate on various hardware platforms. In the following, we discuss selected systems designed to capture biological properties of large-scale neural networks. For this assessment, we only consider communication aspects, e.g. neglecting power consumption.

In the Neurogrid project, the group of K. Boahen developed the design space of their 1 M neuron platform along three axes: the computation of neuron dynamics, the analog or digital implementation style, and the communication architecture [1]. They adopt AER, mapping multiple axons on a single silicon interconnect. Precomputed routing paths are encoded in the package header with optional flooding in the downward traversal of their binary tree. Their deadlock-free routing scheme assures real-time operation with low-precision computations mitigated by population coding.

In the BrainScaleS project, up to 200 k neurons were integrated on a single wafer targeting 10^4x speed-up [23]. High-speed asynchronous AER signaling is realized between clusters of co-located neurons, while pre-configuration eliminates the need for explicit

routing. Each wafer supports 4-5 k neurons when modelling a biological fan-in of SpN without synapse loss. Up to 20 wafers communicate in the system using switched Ethernet network as back-bone.

Adopting a digital design style, the group of S. Furber designed the 1 B neuron SpiNNaker system [7] around a toroidal Mesh6. As a bread and butter algorithm, packets are routed according to LDM with local routing tables supporting redirections and MC. Synchronous AER messages enable modeling of specific axonal delays in contrast to the continuous-time analog signaling. The trade-off between continuous and discrete arrival time is discussed extensively in literature (e. g. [2]).

Recently, Intel's Pohoiki 100 M neuron platform [12] announced supporting biological levels of synaptic densities. An on-chip Mesh4 supports deadlock-free, node-to-node UC, DO routing with hierarchical consolidation on a corresponding chip-to-chip mesh.

5 ASSESSMENT OF PREVAILING COMMUNICATION SCHEMES

To assess the capability of prevailing communication schemes to support the diverse demands in terms of bandwidth and latency, we define three distributions of target neurons N_t:

(1) *uni*: uniform across all nodes,
(2) *rad*: uniform count per radial distance N_t/r_{max} (cf. [18]) within $r_{max} = 8$ mm, and
(3) *bio*: biologically motivated distribution from Section 2.

The first two lend themselves to analytical evaluation, while the latter requires an empirical approach.

5.1 Bandwidth

Fundamentally, the spike generation per node drives bandwidth:

$$B_g = NpN \cdot \nu \cdot a \cdot l_m. \tag{1}$$

For a BC approach, each packet is simply distributed once to all N_t nodes in biological reach:

$$B_{BC} = B_g \cdot N_t. \tag{2}$$

In uni and bio distributions ($N_t = N_N - 1$), this resolves to $B_{BC} \approx 128$ Tbit/s. In the case of UC, each target receives a dedicated packet that travels h_{avg} hops on average:

$$B_{UC} = B_g \cdot SpN \cdot h_{avg}. \tag{3}$$

The derivation of h_{avg} as function of topology follows the method of [10]. For the uni distribution and a Mesh4 topology, for instance, the average hop count computes as $h_{avg} \approx \frac{\sqrt{N_N}}{2}$. This results in a bandwidth requirement for UC of $B_{UC,uni} = 640$ Tbit/s.

Both bandwidth requirements clearly exceed any reasonable technical capability. However, clusters present in the other two distributions can be exploited to reduce the network load. For instance, as the number of possible target neurons in the rad distribution is limited by a cut-off radius, BC messages only need to be transmitted to $N_t = \pi \cdot r_{max}^2 \cdot \rho_N / NpN$ nodes. If this is accounted for in the message distribution scheme, bandwidth requirements are substantially reduced to $B_{BC,rad} \approx 2$ Tbit/s. This plainly highlights the potential of leveraging properties of the biological distribution.

Standing out, MC can ideally provide minimal network load. Stressing the constraint of maximal latency, an exemplary heuristic yielding near optimal results could consist of first computing optimal paths offline using DJ, and then merging overlapping paths. In our case, DJ results in an almost optimal bandwidth reduction because messages from the same node travel on paths which split comparatively late. Since MC bandwidth requirements and bio distributions are hard to assess analytically, we developed a tool that enables empirical calculations of bandwidth and latency requirements [11]. It assumes an isotropic topology, choosing one node as the initial point for calculations and drawing targets randomly. The bandwidth requirements and maximum latency are computed based on the numerical evaluation of number of hops taken by each packet. Its accuracy was validated against the analytical results obtained from Equations 1 – 3.

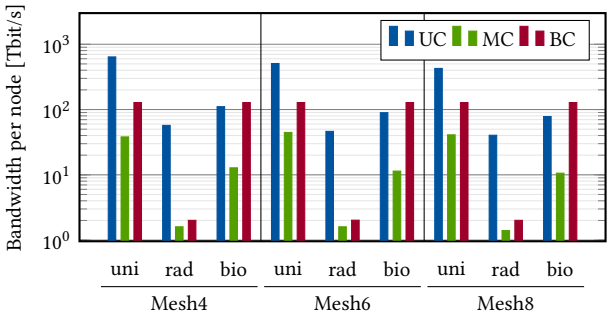

Figure 2: Required link bandwidth and message propagation speed for UC, MC, BC in toroidal mesh topologies.

As depicted in Fig. 2, the bandwidth requirement is significantly affected by the target distribution, with our latency-optimized MC heuristic defining the lower bound. For uni and bio distributions, all schemes exceed the current technical capabilities. However, accounting for r_{max} in a rad distribution improves BC by over 98%, highlighting the advantage of exploiting proximity in clusters.

5.2 Latency

To ensure deterministic results, on time delivery of spike packets is essential. We assume latencies to be constrained by biological distance and axonal speed. As each node represents a square tissue section of a biological neural network with a side length of $\sqrt{\rho_N / NpN}$, this translates to a number of hops each spike has to travel per simulation step, referred to as *speed requirement*.

Assume for example white matter connections in a Mesh8 topology. Then, this translates to a requirement of traversing $v_w \cdot h \cdot \sqrt{\rho_N / NpN} \approx 53$ nodes per time step, i.e. 53 hops/step in this case. Numerical assessments using the aforementioned tool show the speed requirement to span a range from 50 to 75 hops/step for Mesh4, Mesh6 and Mesh8. At the targeted acceleration, a hop latency of less than 20 ns would be required to meet this demand, which is considered technically infeasible with current technology.

6 PROPOSED COMMUNICATION SCHEME

The discussion in the previous section has shown that systems using established communication architectures are by design not capable of reaching the targeted acceleration when accounting for the biologically realistic model from Section 2. Bandwidth requirements of

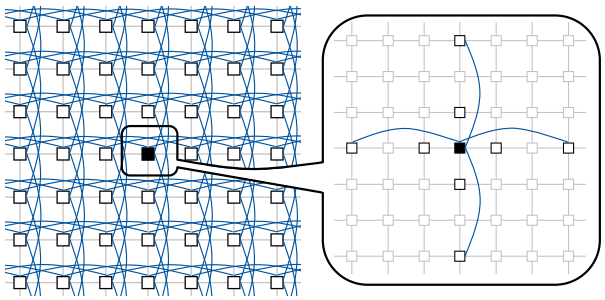

Figure 3: Exemplary construction of a homogeneous Mesh4(1,3) network.

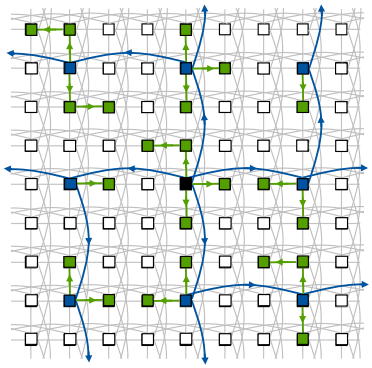

Figure 4: Functionality of BCF in a network with the topology shown in Fig. 3. The first stage (BC) is indicated in blue, the second (UC) in green.

e.g. over 10 Tbit/s for MC in Mesh6 as used in SpiNNaker, and a very low transmission latency of less than 20 ns can hardly be achieved with today's technology. This section introduces a communication scheme designed to address the identified limitations.

Network topology. The challenging latency requirements posed by long-distance high-speed white matter connections can be tackled by adapting the network topology. Starting with a toroidal mesh, the existing connections between direct neighbours must remain to cover small latencies between nearby neurons. We propose to superimpose a homogeneous network of long-distance connections on such a mesh to meet both the low latency requirement of neighbor-to-neighbor communication and the high-speed of white matter connections (see example Mesh4(1,3) in Fig. 3). At the expense of more outgoing links, this structure incorporates shortcuts to nodes further away that contain neurons with tough latency requirements. This also applies to trees and hypercubes. However, both topologies contain adjacent nodes with no direct connections, leading to load imbalances and a higher speed requirement. In contrast, the proposed long hop networks have the same outgoing connections for each node, which additionally facilitates the distribution of messages in the network. Reasonable choices of the range and number of these *long hops* are determined in Section 7.

Casting schemes. The presented long hop connections can help to reduce bandwidth requirements in a network as messages have to be forwarded less frequently. However, this applies only to directed casting schemes such as UC or MC, because the total bandwidth caused by BC is independent of topology and hence much larger than the alternatives, as reasoned in Section 5. MC would be close to optimal in this regard, but the need for offline calculation and routing tables introduces substantial effort we strongly want to avoid for a network size of one billion neurons. Therefore, we consider MC as a reference to which the following alternatives should get as close as possible.

Biologically-realistic connectivity models assume tightly connected clusters in the brain, as described in Section 2. To take advantage of the resulting locality, the distribution of action potential can follow a two stage approach. In the first stage, messages are sent as UC to the center nodes of such clusters. Within their predefined area, messages are then broadcasted in a second stage. This reduces the resulting bandwidth requirements in a similar way to MC. However, the clusters are typically unknown beforehand and need to be located by applying cluster analyses for each source neuron. Furthermore, clusters in white matter connections are not yet well understood. An architecture based on conservative estimates should therefore not take advantage of them.

A more flexible two-stage approach without the need for prior knowledge of cluster locations is the inverse scheme, in the following called BroadcastFirst (BCF) as illustrated in Fig. 4. The first stage is a network BC, exclusively over certain long hop connections with length l. This causes a uniform distribution of messages to single, equidistant nodes in the network that strains only a subset of nodes and links. Since in a homogeneous network each node has identical outgoing long hop connections, the bandwidth load caused by this step is distributed evenly and results in only $1/l^2$ of the total load of a conventional BC. In the second stage, the nodes reached by the BC perform a UC to their $l^2 - 1$ surrounding neighbors which are accessible via very short paths. This step is analogous to the splitting of MC packets which is used in other solutions [7], but does not require special routing tables.

The second approach is applicable as a pure online variant, without any pre-calculations or need for routing tables. It is less efficient for certain distributions when prior knowledge about cluster locations is available, but on the other hand it can efficiently handle a very wide range of different distributions including the worst case scenario of a uniform distribution.

While memory accesses are not within the scope of this work, they pose another important challenge for accelerated neuromorphic simulations. Casting schemes, in which no synaptic information about source and target neuron is contained within the sent messages, require many short random memory accesses to query it. In BCF, each memory entry contains information about the entire neuronal neighbourhood, resulting in significantly fewer accesses. Future work will examine this more closely.

Routing algorithm. The high quality of offline computed routing paths and neuron mappings is only limited by run-time requirements. However, implementing routing tables in the envisioned system poses a major challenge: both, the required capacity and the need for run-time adaption to accommodate plasticity, complicate their design. In contrast, on-the-fly routing algorithms are free of routing tables. Their design is limited by latency and lack of global

oversight, being merely heuristics. Still, implementing lightweight, yet prevailing algorithms like the aforementioned DO and LDM routings might mitigate these downsides. Only small modifications are necessary to have them work with long hop connections and Mesh6 networks. However, since we expect these generic algorithms to not fully exploit long hops, we introduce two additional online routing schemes that inherently support them.

Best neighbour (BN) routing is a lightweight heuristic greedily selecting the link that minimizes the remaining distance to the target. Therefore, it dynamically calculates the distance from each reachable neighbour to the message destination and selects the smallest one. Due to its flexibility, the algorithm is easily applicable without change to systems where the final topology is not yet known or should remain variable. As a less complex alternative, *Longest direction first* (LDR) routing simply chooses the longest outgoing connection that reduces euclidean distance to the target by at least one hop. It is easy to compute since it relies on a static list containing outgoing links sorted by their covered distance.

Synchronization. Late arrival or loss of spikes, which is tolerated in some state-of-the-art systems, would violate the requirement of determinism in the simulation. To prevent this, each node has to wait for all neighbours to finish transmission of relevant spikes. We propose an asynchronous boundary synchronization to ensure a lower bound of propagation speed. More precisely, each node sends a synchronization message to all its neighbors as soon as all spikes are transmitted, which were either generated by the node itself in the current time step or received in the last time step. Redirecting newer messages still continues after synchronization to further accelerate simulation. The number of synchronizations per time step defines the minimum number of hops per time step each message will cover. The actual achieved number of hops per time step can be significantly higher. This local synchronization scheme implicitly decouples areas of nodes, therefore accommodating for spatial and temporal variation in spike rates.

7 EVALUATION

Building on the presented concepts, this section provides an evaluation and refinement towards the 100x acceleration, based on the empirical computation tool mentioned in Section 5. Unless stated otherwise, calculations assume the bio distribution.

Benefit of long hops. To judge the benefit of long hops, we start with three standard topologies (Mesh4/6/8) and expand them with increasing numbers of additional connections. Initial findings showed that lengths of the power of 3 yield reasonably good results (cf. Fig. 5). Thereby, routing paths are determined using MC with DJ to be independent of the routing quality.

Taking the example of Mesh8, the required bandwidth per node drops from 10.6 Tbit/s to 1.3 Tbit/s, and the speed requirement from 50 hops/step to 2 hops/step adding only four long hops. In case of UC, the relative improvement has shown to be even more significant. While requirements decrease with increasing network degree, the additional need for routing resources has to be carefully weighed.

Impact of approximate routing algorithms. In this evaluation, we compared pure Mesh4, Mesh6 and Mesh8 topologies including long hops, and two superpositions of different meshes. These were selected to challenge routing algorithms with different

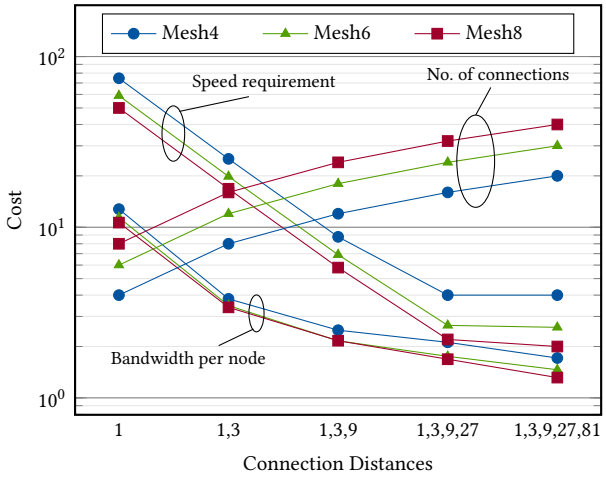

Figure 5: Trade-off of long hops with MC, DJ shown as variation in *cost*, i. e. bandwidth per node (in Tbit/s) or speed requirement (in hops/step) vs. number of connections.

conditions. We started selecting long hops by first defining the longest hop by considering propagation speed in white matter (cf. Section 5), and then shorter hops as fractions thereof. Afterwards, we fine-tuned these to precisely match required speed and improve bandwidth for the most promising routing algorithms. Here, we constrained the connections per node to 32 links as provided by the NetFPGA SUME board.

As reference, basic BC would require 128 Tbit/s independent of topology. A detailed break-down is given in Fig. 6, displaying the bandwidth and speed requirements. The lowest speed requirement we achieved is 2 hops/step.

As expected, DJ provides bandwidth for UC and BCF as a reasonable lower bound. In the case of MC, BN utilizes more shared paths, resulting in lower bandwidth requirements. Overall, LDR proves a reasonable option, while the low-effort methods DO and LDM suffer as they can't adapt direction to exploit long hops. Better results at gradually increased complexity are achieved by BN routing. It is comparable to DJ in all cases, and in a few cases even surpasses it.

Discussion. As expected, MC dominates the solution space due to its global oversight of target connectivity while introducing the aforementioned technical challenges. The next best contender is BCF with BN routing. Both methods excel in the Mesh8(1,3,11,31) topology with a speed requirement of only 2 hops/step.

Let's recall the tremendous improvements of BC requirements in a rad distribution brought by simply limiting message distribution to the biological radius. The bio distribution shares a similar characteristic, having a local centered cluster containing 40% of all targets which is not yet exploited by BCF. As a final optimization step, we therefore introduce BC^2F_l. It extends BCF with an additional local BC limited to a biological distance l.

The quantitative results shown in Fig. 7 highlight the improvements w.r.t. BCF. As expected, MC yields the best results for the rad and clustered bio distributions. The latter requires a bandwidth of only 1.3 Tbit/s which is less than half of BCF. The optimization BC^2F improves bandwidth requirements by approx. 30% compared

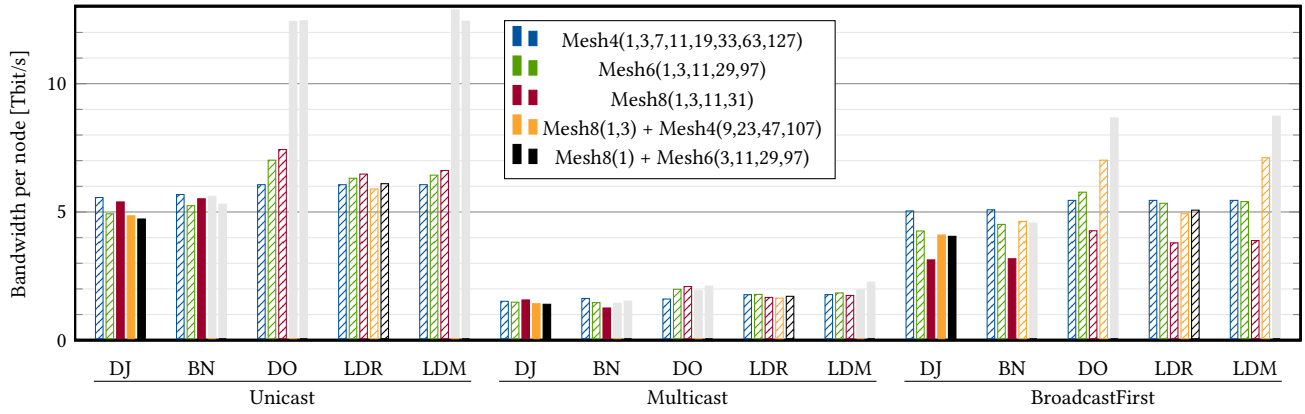

Figure 6: Comparison of routing algorithms for different long hop topologies and bio distribution. Combinations with a speed requirement of 2 hops/step are drawn solid, with a speed requirement > 5 hops/step are grayed out, all in between are shaded.

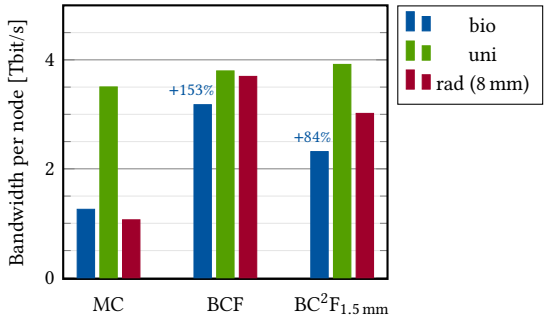

Figure 7: Behavior of MC, BCF and $BC^2F_{1.5\,mm}$ in a Mesh8(1,3,11,31) topology with BN routing (best performing schemes from Fig. 6) for different connection distributions.

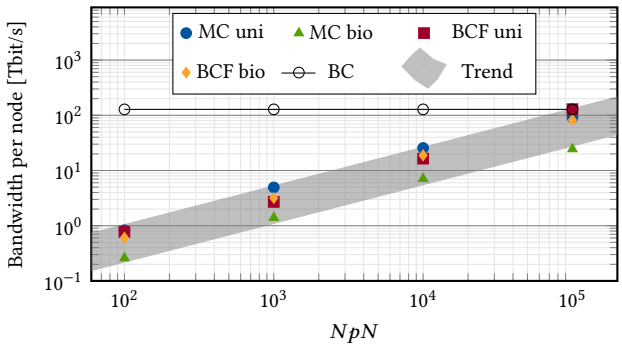

Figure 8: Variation of the number of *NpN* for a constant *N* of 1 G neurons.

to BCF, loading a node with 2.3 Tbit/s and featuring good general capabilities despite its biological motivation. Results for the uni distribution are comparable.

Scaling of *NpN*. So far, we neglected memory accesses and computational effort. As these directly limit NpN, we now consider its effect on bandwidth requirements. Fig. 8 shows a sweep of NpN from 100 to 10^5 indicating a common trend independent of topology and routing algorithm converging towards nodal bandwidth of BC. We conclude that NpN has no significant impact on the quality of the found solution. Furthermore, the sublinear trend indicates that choosing the largest NpN within the mentioned limitations reduces overall system bandwidth.

Dynamic simulation. In order to capture the effect of synchronization and network loading on acceleration, we developed a dynamic simulator. It is written in SystemC, validated against the calculation tool from Section 5 and runs a cycle and bit true model including transceivers, routers and buffers. Fig. 9 presents the achievable acceleration factor as function of one hop latency for incremental changes in the assumptions. All simulations assume a clock frequency of 500 MHz.

When accounting for a speed requirement of 2 hops/step which necessitates 2 synchronization events per time step, acceleration appears as expected slightly below $a \approx h/(2 \cdot hop\ latency)$. Accounting for the need to compute neuron updates, available communication time is halved to 500 ns during simulation of one wall-clock time step of $h/a = 1\,\mu s$. Lastly, the effect of bandwidth limitation is added. Using the theoretical bandwidth requirement per node of 2.3 Tbit/s (BC^2F), the resulting 72 Gbit/s per cable would drop acceleration to 81x. Each of the 32 cables would have to provide 92.5 Gbit/s (+28%) to support 100x acceleration at a node-to-node latency of 100 ns. Conversely, with 50% more bandwidth, an acceleration of 100x can even be reached at 200 ns latency.

8 CONCLUSION

A major challenge in computational neuroscience is the wide-ranging level of detail it addresses, from bio-chemical processes of single ion channels to networks of billions of neurons. This appears familiar to IC design combining nanometer structures with billions of 'active' devices. But neuroscience requires more than off-the-shelf servers and tools due to the brain's intricate structure.

In discussions with neuroscientists, we first gathered their visionary expectations for a neuromorphic simulator and translated

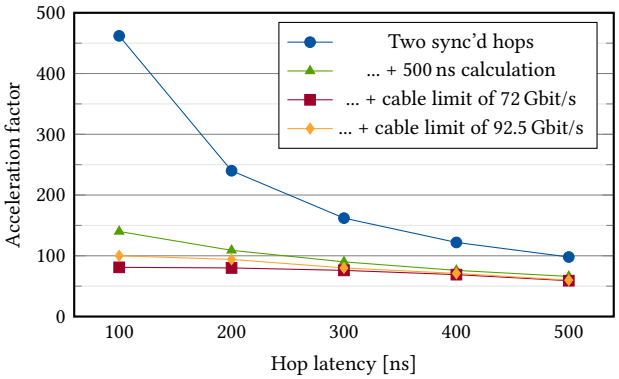

Figure 9: Achievable acceleration factors depending on cable latencies, evaluated in SystemC.

these to a quantitative specification. Applying analytical and numerical techniques, we then showed that prevailing concepts fall short of making the step from today's real-time capable simulators to significantly accelerated simulation of large-scale networks. The results highlighted the importance of adopting a biologically-inspired specification from the very beginning of design.

As a result, we extended the established set of topologies and routing algorithms, and introduced a multi-hop communication topology. Quantitative results, underpinned by bit and cycle accurate simulation, showed that our proposed method brings the envisioned system within the realm of today's off-the-shelf technical communication capabilities. Ultra-low node-to-node latency in the range of 100 ns to 200 ns in combination with nodal bandwidth of up to 4.5 Tbit/s appears sufficient to approach the 100x acceleration with a 0.1 ms resolution in biological time. As these requirements are already met by today's off-the-shelf compute nodes (see Section 3.1), systems supporting the presented network topology can push the acceleration factor of neuroscience simulations at the targeted scale by two orders of magnitude.

ACKNOWLEDGMENTS

This work was partially funded by Helmholtz Society in the Advanced Computing Architectures project (project number SO-092).

REFERENCES

[1] Ben Varkey Benjamin, Peiran Gao, Emmett McQuinn, Swadesh Choudhary, Anand R Chandrasekaran, Jean-Marie Bussat, Rodrigo Alvarez-Icaza, John V Arthur, Paul A Merolla, and Kwabena Boahen. Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations. *Proceedings of the IEEE*, 102(5):699–716, 2014.

[2] Kwabena A Boahen. Point-to-point connectivity between neuromorphic chips using address events. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 47(5):416–434, 2000.

[3] György Buzsáki and Kenji Mizuseki. The log-dynamic brain: how skewed distributions affect network operations. *Nature Reviews Neuroscience*, 15(4):264–278, 2014.

[4] Sergio Davies, Javier Navaridas, Francesco Galluppi, and Steve Furber. Population-based routing in the spinnaker neuromorphic architecture. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2012.

[5] Javier DeFelipe, Lidia Alonso-Nanclares, and Jon I Arellano. Microstructure of the neocortex: comparative aspects. *Journal of neurocytology*, 31(3-5):299–316, 2002.

[6] extoll. High-performance, low latency serdes phy. URL: http://www.extoll.de/products.html, 2020. [Accessed 2020/06/04].

[7] Steve B Furber, Francesco Galluppi, Steve Temple, and Luis A Plana. The SpiNNaker project. *Proceedings of the IEEE*, 102(5):652–665, 2014.

[8] Marc-Oliver Gewaltig and Markus Diesmann. Nest (neural simulation tool). *Scholarpedia*, 2(4):1430, 2007.

[9] Christopher J Glass and Lionel M Ni. The turn model for adaptive routing. *ACM SIGARCH Computer Architecture News*, 20(2):278–287, 1992.

[10] Ellis Horowitz and Alessandro Zorat. The binary tree as an interconnection network: Applications to multiprocessor systems and vlsi. *IEEE Transactions on Computers*, C-30(4):247–253, 1981.

[11] In-house. C++-code. *Will be available via GitLab upon acceptance of paper*, 2020.

[12] Intel. Intel scales neuromorphic research system to 100 million neurons. URL: https://newsroom.intel.com/news/intel-scales-neuromorphic-research-system-100-million-neurons, 2020. [Accessed 2020/06/04].

[13] Jakob Jordan, Tammo Ippen, Moritz Helias, Itaru Kitayama, Mitsuhisa Sato, Jun Igarashi, Markus Diesmann, and Susanne Kunkel. Extremely scalable spiking neuronal network simulation code: from laptops to exascale computers. *Frontiers in neuroinformatics*, 12:2, 2018.

[14] Daniel Liewald, Robert Miller, Nikos Logothetis, Hans-Joachim Wagner, and Almut Schüz. Distribution of axon diameters in cortical white matter: an electron-microscopic study on three human brains and a macaque. *Biological cybernetics*, 108(5):541–557, 2014.

[15] A Theodore Markettos, Paul J Fox, Simon W Moore, and Andrew W Moore. Interconnect for commodity FPGA clusters: Standardized or customized? In *2014 24th International Conference on FPL*, pages 1–8. IEEE, 2014.

[16] Henry Markram. The human brain project. *Scientific American*, 306(6):50–55, 2012.

[17] Pete Mendygral, Nathan Wichmann, Duncan Roweth, Krishna Kandalla, and Kim McMahon. Characterizing full-system network performance and congestion management capabilities with improved network benchmarks. URL: https://cug.org/proceedings/cug2019_proceedings/includes/files/pres125s1.pdf, 2019. [Accessed 2020/06/04].

[18] Javier Navaridas, Mikel Luján, Luis A Plana, Steve Temple, and Steve B Furber. On generating multicast routes for SpiNNaker. In *Proceedings of the 11th ACM Conference on Computing Frontiers*, pages 1–10, 2014.

[19] Umit Y Ogras and Radu Marculescu. " It's a small world after all": NoC performance optimization via long-range link insertion. *IEEE Transactions on VLSI systems*, 14(7):693–706, 2006.

[20] Bente Pakkenberg, Dorte Pelvig, Lisbeth Marner, Mads J Bundgaard, Hans Jørgen G Gundersen, Jens R Nyengaard, and Lisbeth Regeur. Aging and the human neocortex. *Experimental gerontology*, 38(1-2):95–99, 2003.

[21] Tobias C Potjans and Markus Diesmann. The cell-type specific cortical microcircuit: relating structure and activity in a full-scale spiking network model. *Cerebral cortex*, 24(3):785–806, 2014.

[22] Prieto, Dunia and Pérez-Aranda, Rubén. Study of undetected error probability of IEEE 802.3 CRC-32 code for MTTFPA analysis. URL: http://www.ieee802.org/3/bv/public/Jan_2015/perezaranda_3bv_2_0115.pdf, 2015. [Accessed 2020/06/04].

[23] Johannes Schemmel, Johannes Fieres, and Karlheinz Meier. Wafer-scale integration of analog neural networks. In *2008 IEEE IJCNN*, pages 431–438. IEEE, 2008.

[24] Maximilian Schmidt, Rembrandt Bakker, Claus C Hilgetag, Markus Diesmann, and Sacha J van Albada. Multi-scale account of the network structure of macaque visual cortex. *Brain Structure and Function*, 223(3):1409–1435, 2018.

[25] Catherine D Schuman, Thomas E Potok, Robert M Patton, J Douglas Birdwell, Mark E Dean, Garrett S Rose, and James S Plank. A survey of neuromorphic computing and neural networks in hardware. *arXiv preprint: 1705.06963*, 2017.

[26] Shy Shoham, Daniel H O'Connor, and Ronen Segev. How silent is the brain: is there a "dark matter" problem in neuroscience? *Journal of Comparative Physiology A*, 192(8):777–784, 2006.

[27] Armen Stepanyants, Luis M Martinez, Alex S Ferecskó, and Zoltán F Kisvárday. The fractions of short-and long-range connections in the visual cortex. *Proceedings of the National Academy of Sciences*, 106(9):3555–3560, 2009.

[28] Nicole Voges, Almut Schüz, Ad Aertsen, and Stefan Rotter. A modeler's view on the spatial structure of intrinsic horizontal connectivity in the neocortex. *Progress in neurobiology*, 92(3):277–292, 2010.

[29] Samuel S-H Wang, Jennifer R Shultz, Mark J Burish, Kimberly H Harrison, Patrick R Hof, Lex C Towns, Matthew W Wagers, and Krysta D Wyatt. Functional trade-offs in white matter axonal scaling. *Journal of neuroscience*, 28(15):4047–4056, 2008.

[30] SG Waxman and Michael VL Bennett. Relative conduction velocities of small myelinated and non-myelinated fibres in the central nervous system. *Nature New Biology*, 238(85):217–219, 1972.

[31] Stephen G Waxman. Determinants of conduction velocity in myelinated nerve fibers. *Muscle & Nerve*, 3(2):141–150, 1980.

[32] Xilinx. High speed serial transceiver. URL: www.xilinx.com/products/technology/high-speed-serial.html, 2020. [Accessed 2020/06/04].

[33] Noa Zilberman, Yury Audzevich, G Adam Covington, and Andrew W Moore. Netfpga sume: Toward 100 Gbps as research commodity. *IEEE micro*, 34(5):32–41, 2014.