2010 12th International Workshop on Cellular Nanoscale
Networks and their Applications (CNNA)

Scalable Event Routing in Hierarchical Neural Array Architecture with Global Synaptic Connectivity

Siddharth Joshi*, Steve Deiss†‡, Mike Arnold‡§, Jongkil Park*, Theodore Yu*‡, Gert Cauwenberghs†‡
*Dept. of Electrical and Computer Engineering, UC San Diego
†Dept. of Bioengineering, UC San Diego
‡Institute for Neural Computation, UC San Diego
§Computational Neurobiology Laboratory, Salk Institute

***Abstract*—An asynchronous communication scheme for scalable routing of spike events in large-scale neuromorphic hardware is presented. The routing scheme extends the Address-Event Representation (AER) protocol for spike event communication to a modular, hierarchical architecture supporting efficient implementation of global synaptic inter-connectivity across a cellular matrix of message parsing axonal relay nodes at varying spatial scales. This paper presents a probabilistic framework for analyzing trade-offs in throughput and latency of synaptic communication as a function of load and geometry, and simulation results verifying the statistics of traffic flow across the architecture.**

I. Introduction

One of the major challenges faced in neuromorphic engineering is the vast numbers of neurons and synapses involved in modeling cognitive neural systems, such as cortical models of visual information processing. The combination of locally dense and sparsely global synaptic connectivity makes it virtually impossible to implement such systems with dedicated wiring for the synaptic connections using conventional 2-D silicon microtechnology. The need for 'virtual wires' in implementing global forms of connectivity between multiple microchips implementing neural functions has prompted research into efficient communication between multiple neuromorphic chips.

Early research in multichip communication between neuromorphic chips led to the Address-Event Representation (AER) protocol [1]. AER was originally formulated for time division multiplexing of events to emulate the high degree of connectivity between neurons [1]–[3], and further extended with specialized addressing schemes to ensure on-time delivery of messages to their destinations [4]. Over time AER became the *de facto* protocol for spike based inter-chip communication to connect neurons across multiple chips [5]–[7]. The combination of dense memories, digital logic, and analog neurons enabled the implementation of large scale reconfigurable neural architectures that extend on the basic AER protocol to achieve greater functionality in modeling synaptic connectivity and plasticity [8], [9], such as spike timing-dependent plasticity in the address domain [10], [11].

While the extended AER synaptic routing protocol provides, in principle, for unlimited spatial connectivity between neurons, bandwidth limitations in a single-bus implementation of AER constrain the size and activity level of the network, where the product of neural activity and synaptic connectivity cannot exceed this bandwidth. To overcome this limitation, grid based approaches have been proposed, where multiple AER routing nodes are connected in a mesh that operates in systolic fashion [12]. Although these multi-bus grid approaches improve on the bandwidth limitations of single-bus AER, limitations on memory use and latencies in the implementation of typical large-scale neural systems, due to the wide range in spatial synaptic connectivity and temporal dynamics in axonal propagation, are challenges for scalable implementation that are addressed in this paper.

We extend AER spike communication from a flat architecture to a fractal hierarchy. Specifically we extend flat AER, with a single address bus shared [8], [9] or a grid of address routers distributed [12] across neurons, to hierarchical AER, with repeated address buses and communication relays at varying spatial scales of synaptic interconnectivity. We partition delays and provide provisioning for a hierarchical delay scheme. This extension is critical in scaling up neuromorphic systems towards levels of synaptic connectivity approaching that of the human central nervous system. We present a probabilistic framework to quantify throughput and latency of synaptic communication as a function of load and geometry, and verify these findings with simulations on statistics of event traffic and queue occupancy for different geometric spread of synaptic fan-out.

II. Architecture

The proposed synaptic routing architecture (Fig. 1) is based on the address-event representation (AER), but differs from conventional [1] and grid [12] AER routing due to the multi-scale hierarchical, rather than flat representation of address events, as well as the partitioning of event delays in a hierarchically optimal manner. Its ability to efficiently route events through global hierarchies, and to model the wide range of

978-1-4244-6678-8/10/$26.00 ©2010 IEEE

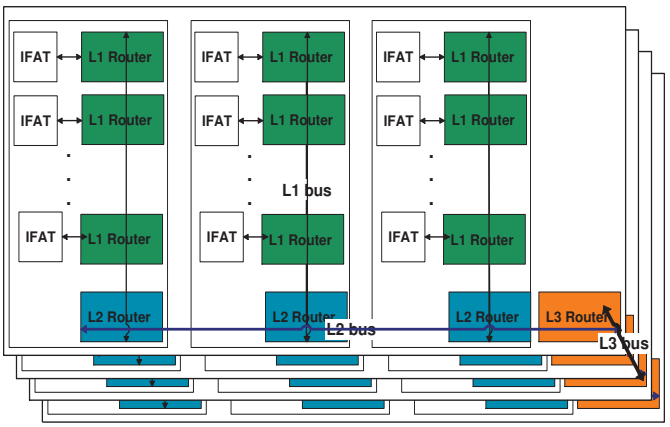

Fig. 1. Multi-chip hierarchical spike event communication architecture, with integrate-and-fire array transceiver (IFAT) neuronal blocks, and multi-level neural event routers

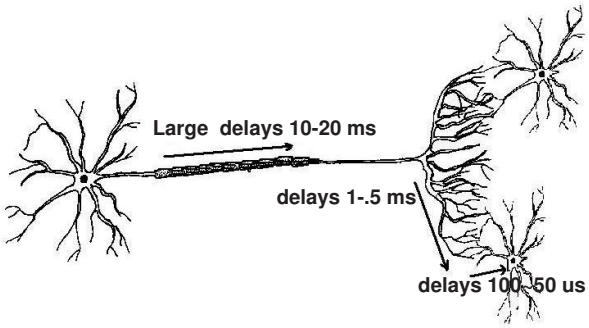

Fig. 2. Topological event grouping and partitioning of axonal delays. Synaptic events with larger delays, traveling larger distances, are grouped according to presynaptic source addresses, thereby reducing queue occupancy (Sec. III-C).

delays in axonal and dendritic propagation through such global structures, presents a significant advantage.

A. Multi-Level Synaptic Event Routing

In the proposed architecture, synaptic events are pooled by presynaptic address, for efficient routing to destination areas at multiple levels of increasing spatial scale. The sequence of routing steps from a presynaptic source event to multiple postsynaptic destination entails the following:

1) The presynaptic neuron sends a single event locally.
2) Each routing node recognizes the event by a local address, unique to the presynaptic neuron and routes the event to parents, children and/or siblings as needed, based on the connectivity at that level of spatial scale. It utilizes local storage as an address look-up table to identify the immediate destinations of the event, where this process is repeated. All along, the propagating messages represent the same presynaptic event throughout the hierarchy.
3) In routing hops, logarithmic in the size of the network, the presynaptic event is able to reach the leaf nodes in the hierarchy (at the postsynaptic neuron level). At each destination, the local address of the event indexes the synaptic table for the local postsynaptic addresses and synaptic parameters. This is the only place in the entire hierarchy where postsynaptic data is represented.

By virtue of source encoding, events are bundled from one source to multiple destinations into one transmitted message, hence more effectively making use of the available bandwidth. It should be noted though, that using this scheme means that the routing elements need *a priori* information on the destinations for each source. In a flat system, this requires that each source have a unique global address, severely reducing the total number of neurons the system can accommodate. In the hierarchical system, only the next destination for the original source needs to be stored at each routing node, thus aliasing is not a factor. Another feature of the hierarchical partitioning is that neurons that are spatially separated in the biological system can be mapped to spatially separated nodes in the architecture. Since these nodes communicate less frequently than those closer together, the bandwidth for communicating between chips is not taxed as heavily as that within the chip, which in turn is not taxed as heavily as the bandwidth within a neuronal cell/array.

The hierarchical partitioning of the system also enables the reliable implementation of delays consistent with those found within a biological system [11]. Within a biological system axonal delays and their contribution to learning have been widely studied, they are modeled as propagation delays as shown in Fig. 2. Similarly in this architecture, the source and destination of messages preserve the topology of their counterparts in biological systems. The delays are implemented as waiting times within queues. A combination of large and fine delays results in a high degree of control. Each level of the hierarchy implements delays at a temporal resolution that is progressively coarser with the distance from the lowest level. Because of the pooling of longer range, slower synaptic events according to presynaptic address, fewer events are routed at the higher levels, thereby reducing the occupancy of long wait events in the router queues.

B. Implementation

Figure 1 provides an overview of the simulated architecture. The example given assumes integrate-and-fire array transceiver (IFAT) based neurons, but the architecture is independent of the details of neuronal implementation. The neurons are physically present in a neuronal array at the lowest level (L0) region, along with the synaptic look-up tables. When an event arrives the L0 arbiter, through a series of synaptic fan-out tables (FOT), looks up the addresses of the synaptic recipients. If the recipient is physically located within this L0 array of neurons, the event is looped back, with required adjustments made in the weights of the synapses. If the destination is in a different address space, the source information is placed on the level-1 (L1) bus, and a message copy is sent to each sibling that is part of the destination. If the destination is off-chip, the same sequence of routing events is replicated at a larger, multi-chip scale. This form of event-messaging can be replicated at

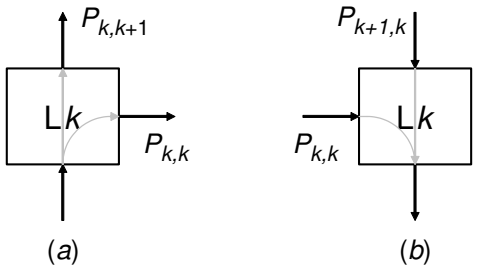

Fig. 3. Probabilistic model of event routing at level *k* in the routing hierarchy. *(a)*: Event messages from children are routed to siblings and the parent. *(b)*: Event messages from parents or siblings are routed only to children.

each level of hierarchy as described below. A sample router at any level-k (Lk) is shown in Fig. 3, and assumes a probabilistic model of the following communication scheme:

1) Any node can communicate either to its siblings or to its parents if it has to send a message received from its child (trickle up).
2) If a message is received by a node from its siblings or its parents it will send those events only to its children (trickle down).
3) In case the same message has to be sent to multiple children or multiple siblings, that message will be replicated.

In the implementation, all arbiters/routers are further equipped with FIFO structures into which they accept events, process them, and 'pop' them. The FIFO processing time contributes to the overall delay experienced by an event and together with the modeled axonal delays determines the event rate capacity of the routing network as well as the queue occupancy at the routing nodes.

III. Analysis

In what follows we derive and analyze, using physically based assumptions and approximations, expressions for global and local fan-out, global event traffic, and queue occupancy as a function of geometric parameters related to connection topology and propagation delays, inspired by the spatial distribution and delay properties of axons and white matter. The analysis follows general principles of routing in fat tree-based network architecture for which extensive theory and analysis exists [13], which applies directly to the hierarchical tree-based AER framework presented above.

A. Global and local synaptic fan-out

Let the global synaptic fan-out be be given by S, and the local synaptic fan-out, originating from level i in the hierarchy (Li) be given by S_i. We also define $P_{i,i+1}, P_{i,i}$ and $P_{i,i-1}$ as the probabilities of parent, sibling, and child event transmission, respectively from level i to parent at level $i+1$, from one node to all its siblings at level i, and from a parent at level i to its child at level $i-1$. We further assume a tree-based topology of the routing hierarchy with a branching factor n between consecutive levels i and $i+1$, and with depth d, $i = 0, 1, \ldots d$. We then probabilistically express the global synaptic fan-out in terms of the local synaptic fan-outs at various levels i as:

$$\begin{aligned} S &= S_0 + (n-1)P_{1,1}S_1 \\ &\quad + P_{1,2}(n-1)P_{2,2}nP_{2,1}S_2 \\ &\quad + P_{1,2}P_{2,3}(n-1)P_{3,3}nP_{3,2}nP_{2,1}S_3 + \ldots \end{aligned} \tag{1}$$

In the special case of a flat hierarchy, we may assume a uniform connectivity at multiple scales, *i.e.,* $S_0 = S_1 = \ldots = S_d = S_\ell$. Therefore, for a hierarchy of depth d the global connectivity S is given by $n^d S_\ell$, or NS_ℓ where N is the number of leaf nodes in the hierarchy. In other words, each of the N L0 routers shares an equal fraction S/N of the global connectivity bandwidth S. More generally, for various degrees of locality in the connectivity, we define a geometric spread parameter $0 \leq \lambda \leq 1$ that approximates the event transmission probabilities across the hierarchy of spatial connectivity scales in a geometrical series:

$$\begin{aligned} S_i &= S_\ell \\ P_{i-1,i} &= 1 \\ P_{i,i} &= \lambda \\ P_{i,i-1} &= \lambda \end{aligned} \tag{2}$$

expressing a fractal geometry pattern in the synaptic connectivity, with equal branching of fan-out at each spatial scale. This leads to a geometric power series for the fan-out (1):

$$\begin{aligned} S &= (1 + (n-1)\lambda + (n-1)n\lambda^2 + (n-1)n^{k-1}\lambda^k)S_\ell \\ &= \frac{1 - \lambda - (n-1)n^d\lambda^{d+1}}{1 - n\lambda} S_\ell \end{aligned} \tag{3}$$

in the branching factor n, geometry spread λ, and up to a depth d of the hierarchy. Note that this series converges to a finite global connectivity at infinite depth, $\lim_{d\to\infty} S = (1-\lambda)/(1-n\lambda)\ S_\ell$, for a spread parameter smaller than the branching ratio, $\lambda < 1/n$. In general, for finite depth d, we consider values of spread λ between 0 and 1, corresponding to the extremes of strictly local and strictly global synaptic connectivity, respectively. The dependence between global S and local S_ℓ synaptic connectivity, for various branching factor n and geometry spread λ, is illustrated in Fig. 4.

B. Event global messaging traffic through the hierarchy

The traffic of global messages throughout the hierarchy, in service to a neural event at any local source location,

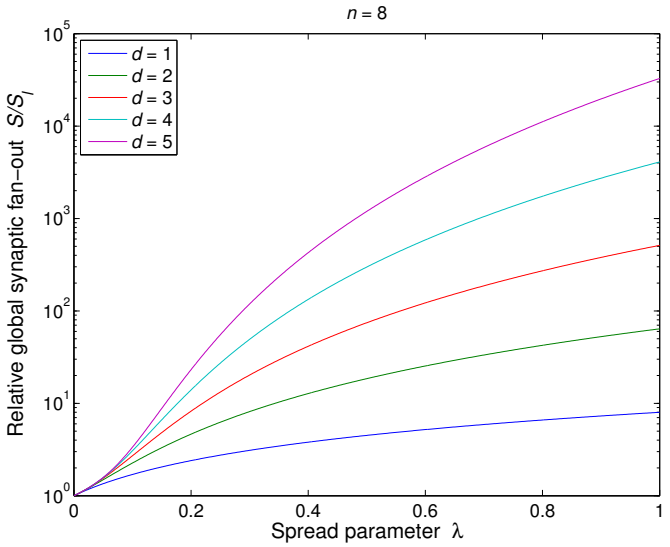

Fig. 4. Global synaptic fan-out S relative to local fan-out S_l as a function of geometric spread factor λ.

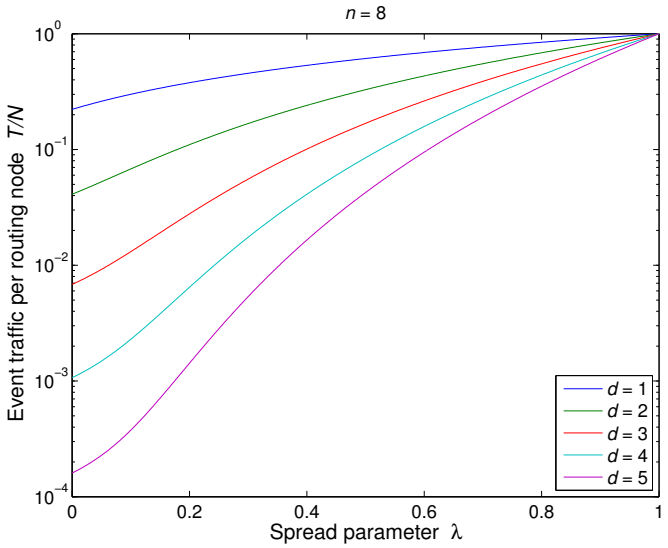

Fig. 5. Traffic overhead T/N, as the number of global messages per spike event per routing node, as a function of geometric spread factor λ.

decomposes at different levels of synaptic reach Lk as follows:

$$\begin{aligned}
\mathrm{L0}: &\quad 1+S_0\\
\mathrm{L1}: &\quad 1+(n-1)P_{1,1}+(n-1)P_{1,1}S_1\\
\mathrm{L2}: &\quad P_{1,2}+P_{1,2}(n-1)P_{2,2}\\
&\quad +P_{1,2}(n-1)P_{2,2}nP_{2,1}+P_{1,2}(n-1)P_{2,2}nP_{2,1}S_2\\
\mathrm{L}k: &\quad P_{1,2}P_{2,3}\dots P_{k-1,k} \qquad (4)\\
&\quad +(n-1)P_{1,2}P_{2,3}\dots P_{k-1,k}P_{k,k}\\
&\quad +(n-1)nP_{1,2}P_{2,3}\dots P_{k-1,k}P_{k,k}P_{k,k-1}+\dots\\
&\quad +(n-1)n^{k-1}P_{1,2}P_{2,3}\dots P_{k,k}P_{k,k-1}\dots P_{3,2}P_{2,1}\\
&\quad +(n-1)n^{k-1}P_{1,2}P_{2,3}\dots P_{k,k}P_{k,k-1}\dots P_{3,2}P_{2,1}S_k.
\end{aligned}$$

In the special case of strictly local connectivity, only the first term is non-zero, and all messaging is strictly local. The other extreme case of interest is a flat globally connected hierarchy ($lambda = 1$), $P_{i,j} = 1$ and $S_i = S_\ell$, for which the next messaging traffic across the hierarchy is given by $1+n+n^2+\dots+n^d+n^dS_\ell = (1-n^{d+1})/(1-n)+n^dS_\ell$. If $S_\ell \gg 1$ or $S \gg n^d$, then the traffic overhead of a flat implementation over a hierarchical implementation is negligible. In general, for $0 \le \lambda \le 1$ we consider $S_i = S_\ell, P_{i,i} = P_{i,i-1} = \lambda, P_{i,i+1} = 1$ as above, resulting in a total traffic $S+T$, where S is the net combined local synaptic messaging at all postsynaptic sites in the hierarchy, and where T is the total overhead traffic in combined global messaging to deliver the presynaptic source event to each of its postsynaptic destinations:

$$T = d+1+(n-1)\lambda\left(\frac{d}{1-n\lambda}-\frac{n\lambda(1-(n\lambda)^d)}{(1-n\lambda)^2}\right). \qquad (5)$$

It is interesting to relate this net messaging overhead in global traffic to the number of routing nodes in the hierarchy:

$$N = 1+n+n^2+\dots n^d = \frac{1-n^{d+1}}{1-n}. \qquad (6)$$

For a global uniform connectivity $\lambda = 1$, this results in a worst-case traffic overhead per router T/N of 1. In other words, in the worst case of full global connectivity, each router in the hierarchy sees each neural event passing by just once, regardless of the synaptic fan-out. In the other extreme of strictly local connectivity $\lambda = 0$, the traffic overhead is zero, $T/N = 0$, and all event messaging is local to serve the local postsynaptic sites. The general case $0 \le \lambda \le 1$ falls between these extremes with $0 \le T/N \le 1$, illustrated in Fig. 5.

C. Delay partitioning for reduced queue occupancy

To analyze queue occupancy at each of the routing nodes due to the required wait times imposed by the modeled axonal delays, we further consider a geometrically structured dependence of axonal delays organized along the routing hierarchy, as illustrated in Fig. 2. We denote τ_i as the average total axonal (and dendritic) delay of a connection between a presynaptic and postsynaptic neuron crossing at level k in the hierarchy, where we assume that typically these delays increase with increasing spatial scale: $\tau_i > \tau_j$ for $i > j$. For efficient routing of these events (and also modeling the geometric bundling of event transmission along axonal fiber bundles in white matter), we take advantage of the typically increased sparsity of event messaging at larger spatial scales. In particular, we partition the axonal (and dendritic) delays τ_i into queue wait times distributed across the routers in the path of the synaptic transmission:

$$\begin{aligned}
\tau_i \quad = \quad & \tau_{0,1}+\tau_{1,2}+\dots+\tau_{i-1,i}\\
& +\tau_{i,i}+\tau_{i,i-1}+\tau_{i-1,i-2}+\dots+\tau_{1,0} \qquad (7)
\end{aligned}$$

where $\tau_{i,i}$ is the delay (queue wait time) in the event routing at level i to siblings, $\tau_{i,i+1}$ is the delay in the event routing from a child at i to its parent at $i+1$ (0), and $\tau_{i,i-1}$ is the event routing from a parent at i to its children at $i-1$.

Without loss of generality, and to minimize timing errors in the implementation, we may assume $\tau_{i,i+1} = 0$ since an event does not multiply during up-level transmission, $P_{i,i+1} = 1$ with unity event fan-out.

The global queue occupancy Q_{hier} for synaptic events at the level i connectivity scale, for an event rate ν per L0 routing node (spike rate times the number of local neurons) and corresponding propagation of these events through the cascade of parents, siblings, and children, is then quantified by summing the individual queue occupancies approximated by Little's law [14] in each of the routing nodes in the synaptic path as:

$$\begin{aligned} \mathrm{L}i: &\quad \nu\tau_{i,i}P_{1,2}P_{2,3}\dots P_{i-1,i}(n-1)P_{i,i} \\ \mathrm{L}i-1: &\quad \nu\tau_{i,i}P_{1,2}P_{2,3}\dots P_{i-1,i}(n-1)P_{i,i} \\ &\vdots \\ \mathrm{L}1: &\quad \nu\tau_{2,1}P_{1,2}P_{2,3}\dots P_{i-1,i}(n-1)P_{i,i} \\ &\quad nP_{i,i-1}nP_{i-1,i-2}\dots nP_{2,1} \\ \mathrm{L}0: &\quad \nu\tau_{2,1}P_{1,2}P_{2,3}\dots P_{i-1,i}(n-1)P_{i,i} \\ &\quad nP_{i,i-1}nP_{i-1,i-2}\dots nP_{2,1}S_i. \end{aligned} \tag{8}$$

In contrast, for a flat hierarchy without intermediate routing nodes and without partitioning of the delay τ_i (7), the flat queue occupancy Q_{flat} for synaptic events at the level i connectivity scale subject to the τ_i wait time is given by

$$\begin{aligned} &\nu\tau_i\dots P_{1,2}P_{2,3}\dots P_{i-1,i}(n-1)P_{i,i} \\ &nP_{i,i-1}nP_{i-1,i-2}\dots nP_{2,1}S_i. \end{aligned} \tag{9}$$

The relative queue occupancy, for the hierarchical vs. the flat routing architecture, is then given by

$$\frac{Q_{hier}}{Q_{flat}} = \frac{\frac{\tau_{1,0}}{1} + \frac{\tau_{2,1}}{\sigma_1 S_i} + \frac{\tau_{3,2}}{\sigma_2 S_i} + \dots + \frac{\tau_{i,i-1}}{\sigma_{i-1} S_i} + \frac{\tau_{i,i}}{\sigma_i S_i}}{\tau_{1,0} + \tau_{2,1} + \tau_{3,2} + \dots + \tau_{i,i-1} + \tau_{i,i}} \tag{10}$$

where

$$\sigma_j = n^{j-1}P_{j,j-1}\dots P_{3,2}P_{2,1}. \tag{11}$$

Q_{hier} is thus guaranteed lower than Q_{flat}, leading to savings in memory hardware resources, when $\sigma_j S_i > 1$ for all $j \le i$. For large local synaptic fan-out S_ℓ this is usually the case. More generally, the delay distribution also plays an important role in relative queue occupancy for hierarchically partitioned delay distribution. To this end we introduce a geometric stall parameter μ, which quantifies the degree of progressive slow-down in the partitioned delays at increasing spatial scale. Specifically, we assume another geometric series in spatial scale i, now for the axonal/dendritic delays τ_i with stall parameter μ:

$$\begin{aligned} \tau_{j,j+1} &= 0 \\ \tau_{j+1,j} &= \tau_0\mu^j \\ \tau_{i,i} &= \tau_0\mu^i \end{aligned} \tag{12}$$

for all $0 \le j \le i-1$. Together with the geometric dependence of event messaging through the spread parameter λ, the geometric stall dependence (12) yields the following expression

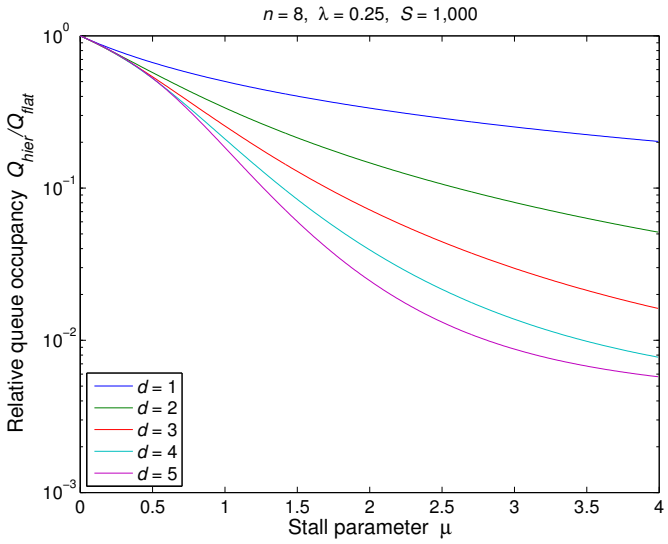

Fig. 6. Queue occupancy in hierarchical routing Q_{hier} relative to flat synaptic routing Q_{flat}, as a function of geometric stall factor μ.

for the relative queue occupancy (10):

$$\frac{Q_{hier}}{Q_{flat}} = \left(1 + \frac{\mu}{S_\ell}\frac{1-(\frac{\mu}{n\lambda})^i}{1-\frac{\mu}{n\lambda}}\right) \Big/ \left(\frac{\mu^{i+1}-1}{\mu-1}\right) \tag{13}$$

In the pathological case of zero stall $\mu = 0$, all delay is concentrated at the postsynaptic level, and no savings result, $Q_{hier}/Q_{flat} = 1$. The more realistic and interesting case $\mu > 1$, with increasing delays at increasing spatial scales, may yield significant savings in queue occupancy for large S_ℓ, with guaranteed savings for $\mu \ge n\lambda$ (although this condition is not essential). Example curves for this dependency as a function of stall μ and depth d (i) are shown in Figure 6.

IV. Experiments and Results

We created a model of the message passing behavior in this network along with realistic assumptions on traffic load at each level of hierarchy assuming a uniform, randomly connected network. Results from simulating the network using a hierarchical implementation with depth $d = 1$, in comparison with those obtained by simulating the equivalent network in a flat representation $d = 0$, are shown in Fig. 7.

The product of the average delay faced by an element τ with the rate of incoming events ν as well as the average of the number of events in the queue Q, are displayed. According to Little's law [14] $\nu\tau = Q_{ave}$ where Q_{ave} is the long time average of elements in the queue. The rate of incoming events ν is essentially the number of neurons, local at the L0 level, multiplied by the average rate of fire for each neuron. This queue can be treated as a system in queuing theory and analyzed using established results from queuing theory and network theory. The results can be used for creating an efficient design and deriving performance bounds. Queuing theory based models have been very well studied with multiple characteristics, thus their use allows for great flexibility in analyzing a large variety of implementations [15].

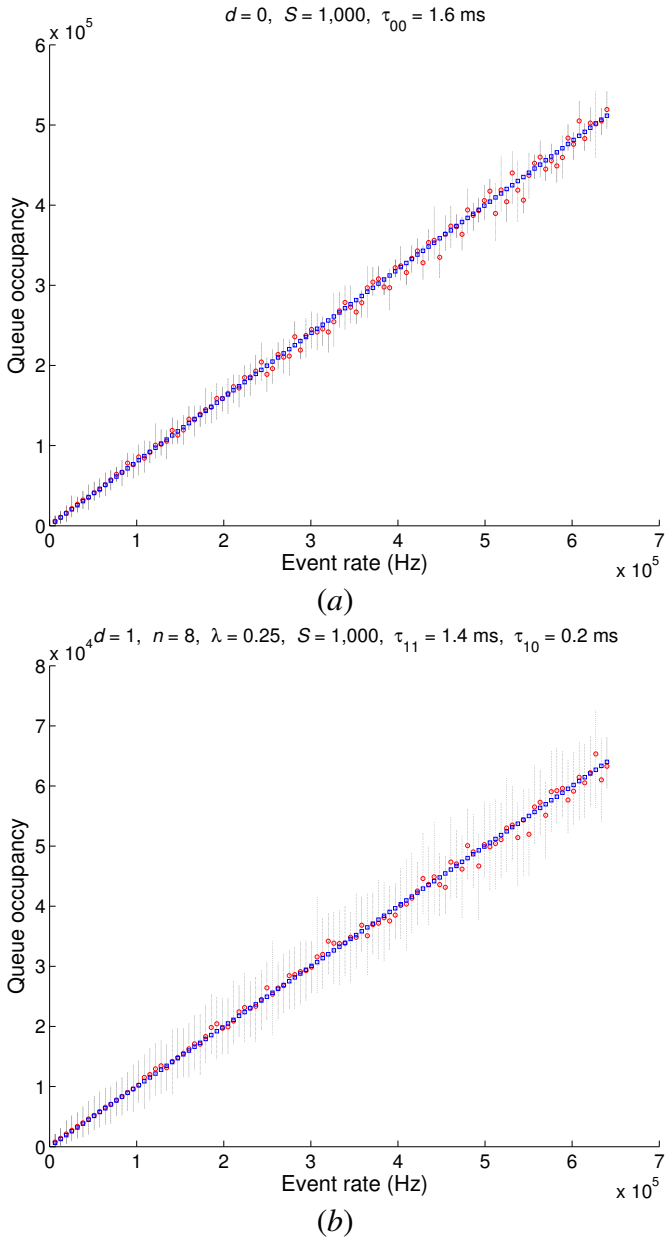

Fig. 7. Observed random, average, and theoretical queue occupancy as a function of event rate, (a) for a flat hierarchy with delay $\tau_0 = 1.6$ ms, and (b) for a depth $d = 1$ hierarchy with $\lambda = 0.25$, $n = 8$, $\tau_{11} = 1.4$ ms, and $\tau_{10} = 0.2$ ms. Global synaptic fan-out $S = 1,000$ in both cases.

V. CONCLUSIONS

In this paper a modular architecture for scalable, hierarchical AER asynchronous spike event routing was presented. By virtue of fractal hierarchy, spike events may be efficiently communicated to neurons over a wide distribution of distances, supporting global synaptic inter-connectivity with nearest-neighbor cellular communication at varying spatial scales. We presented trade-offs between throughput and latency of spike communication as a function of synaptic density, spike rate, and FIFO buffer length, and verified these findings with statistical results obtained from simulations on a small-scale version of the architecture. We analyze the presented architecture with respect to queuing theory and show a close correspondence between the results. This results in an ease in evaluating trade-offs for designing the hardware system. These simulation results indicate that a synaptic fan-out of 1,000 can be sustained across a network of arbitrary large scale, implemented in parallel assuming today's standard processing and memory limits. Future work is directed towards integrating the implementation of the asynchronous communication architecture with spiking neuromorphic analog hardware.

REFERENCES

[1] M. Mahowald, *An analog VLSI system for stereoscopic vision*. Boston: Kluwer Academic Publishers, 1994.

[2] K. Boahen, "A throughput-on-demand address-event transmitter for neuromorphic chips," in *ARVLSI*, 1999, pp. 72–87.

[3] E. Culurciello, R. Etienne-Cummings, and K. Boahen, "High dynamic range, arbitrated address event representation digital imager," in *ISCAS (3)*, 2001, pp. 505–508.

[4] S. Deiss, "Connectionism without the connections," in *Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on*, vol. 2, Jun-2 Jul 1994, pp. 1217–1221 vol.2.

[5] S. Deiss, R. Douglas, and A. Whatley, "A pulse-coded communications infrastructure for neuromorphic systems," *Pulsed Neural Networks, MIT Press*, (Mass W., Bishop, C.M., ed), 1999.

[6] M. Khan, D. Lester, L. Plana, A. Rast, X. Jin, E. Painkras, and S. Furber, "Spinnaker: Mapping neural networks onto a massively-parallel chip multiprocessor," in *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, June 2008, pp. 2849–2856.

[7] J. Fieres, J. Schemmel, and K. Meier, "Realizing biological spiking network models in a configurable wafer-scale hardware system," in *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, June 2008, pp. 969–976.

[8] R. Vogelstein, U. Mallik, J. Vogelstein, and G. Cauwenberghs, "Dynamically reconfigurable silicon array of spiking neurons with conductance-based synapses," *Neural Networks, IEEE Transactions on*, vol. 18, no. 1, pp. 253–265, Jan. 2007.

[9] R. Serrano-Gotarredona, M. Oster, P. Lichtsteiner, A. Linares-Barranco, R. Paz-Vicente, F. Gomez-Rodriguez, K. Camunas-Mesa, R. Berner, M. Rivas-Perez, T. Delbruck, S. Liu, R. Douglas, P. Hafliger, G. Jimenez-Moreno, A. Ballcels, T. Serrano-Gotarredona, A. Acosta-Jimenez, , and B. Linares-Barranco, "CAVIAR: A 45k neuron, 5M synapse, 12G connects/s AER hardware sensory-processing-learning-actuating system for high-speed visual object recognition and tracking," vol. 20, no. 9, 2009, pp. 1417–1438.

[10] R. J. Vogelstein, F. Tenore, R. Philipp, M. S. Adlerstein, D. H. Goldberg, and G. Cauwenberghs, "Spike timing-dependent plasticity in the address domain," in *NIPS*, 2002, pp. 1147–1154.

[11] J. V. Arthur and K. Boahen, "Learning in silicon: Timing is everything," in *NIPS*, 2005.

[12] P. A. Merolla, J. V. Arthur, B. E. Shi, and K. A. Boahen, "Expandable networks for neuromorphic chips," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 54, no. 2, pp. 301–311, Feb. 2007.

[13] C. E. Leiserson, "Fat-trees: universal networks for hardware-efficient supercomputing," *IEEE Trans. Comput.*, vol. 34, no. 10, pp. 892–901, 1985.

[14] J. D. C. Little, "A proof for the queuing formula: $l = \lambda w$," *Operations Research*, vol. 9, no. 3, pp. 383–387, 1961.

[15] E. Gelenbe and G. Pujolle, *Introduction to queueing networks*. New York, NY, USA: John Wiley & Sons, Inc., 1987.