Energy Constraints for Building Large-Scale Systems

Jennifer Hasler
Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA, USA

Abstract: *Computational efficiency discussions are necessary for understanding how to build an energy efficient cortical structure, but not sufficient because we need to consider the resulting power dissipation for communication. Neurobiological systems are power (and energy) constrained in their communication. Any cortical architecture must explicitly incorporate these effects to achieve the necessary power efficiency gains, although most systems built to date do not consider these issues as primary constraints.*

Keywords: Neuromorphic Engineering; Cortical Operation, Dendritic Computation; FPAAs; Floating-Gate

Computational efficiency (computation per unit energy) considerations are necessary for understanding how to build an energy-efficient cortical structure to reach equivalent computational levels of 1-100MMAC(/s)/nW in Biological and Si neurobiological levels [1], but not sufficient because we need to consider the resulting power dissipation for communication [1]. Neurobiological systems are power (and energy) constrained in their communication. The human cortex consumes about 20W of power, of which, only a fraction (< 25%) of this power is used for communication, limiting the number and length of axonal connections. Most biological neurons have a high level of local interconnection, particularly cortical neurons.

Any cortical architecture must explicitly incorporate these effects to achieve the necessary power efficiency gains, although most systems built to date do not consider these issues as primary constraints. The DARPA Synapse program, requiring very tight constraints in other areas (e.g. Synapses), basically ignored any realistic power constraint for their resulting architectures allowing kW of power for a mouse brain verses 20W actually consumed by the human brain. .

Therefore, most of the computation needs to be local; fortunately, neurobiological systems use a similar approach in the fact that over 90% of neurons in cortex project locally to nearby neurons (i.e. nearest 1000 pyramidal cells). We want to have as much communication locally on a single IC for low-power operation. Integrating memory and computation, as in biological systems also keeps communication power manageable. Using external memory as the primary approach for programmability and configurability, as is the typical use of Address-Event Representation (AER) communication schemes, comes at a huge cost that makes scaling to large systems impractical.

Constraints from Biological Computation

Computational efficiency discussions are necessary for understanding how to build an energy efficient cortical structure, but not sufficient because we need to consider the resulting power dissipation for communication. We find that neurobiological systems are constrained in their communication because of power constraints [1]. The human cortex consumes about 20W of power, of which, only a fraction (25%) of this power is used for communication, limiting the number and length of axonal connections. This result is consistent with data that most neurons have a high level of local interconnection [2], such as nearby cortical neurons; any cortical architecture must explicitly incorporate these effects to achieve the necessary power efficiency gains. The result requires most of the computation to be local; fortunately, neurobiological systems use a similar approach in the fact that over 90% of neurons in cortex project locally to nearby neurons (i.e. nearest 1000 pyramidal cells).

Neurons primarily communicate to other neurons primarily communicating events, or action potentials, which are effectively digital signals. These digital events can be modeled similarly to Si digital communication down a transmission line, where energy is proportional to the capacitive load, and quadratically dependant on the power supply (V_{dd}). V_{dd} for a biological communication is between 100mW to 180mW [3]. Given the power consumed per neuron output with a typical cortical event rate (0.5 Hz firing rate) results in roughly 250pF total capacitance on an axon line for a biological system, corresponding to 30.6mm average total cable length of 1μm diameter axon cable (fairly thin axon; typical axonal diameters are 1μm to 20μm). Considering mylenation for cortical axons only slightly (3-5%) changes this total cable length.

The net result is that with most communication on biological axon lines, even though they might be present everywhere, including intricate three-dimensional patterns, one does find an exponentially decreasing distribution of axon cable length in cortex, consistent with the neural communication being constrained to a tight power budget. This result is consistent with data that most neurons have a high level of local interconnection [2], such as nearby cortical neurons; any cortical architecture must explicitly incorporate these effects to achieve the necessary power efficiency gains. Further, these results are also consistent with the low average spike rates found in cortical systems

(1 spike per second); an entire cortical network operating with rate encoded signals (i.e. 3 to 300 Hz) would consume 100 times the power, and therefore the axon cable length for a cortical power dissipation requires 100 times shorter cables, which is impractical. We expect that constraining silicon communication power may be required based on this biological inspiration.

Constraints from Digital Computation Systems

Rarely is the digital communication included in power for computation, although often it is the limiting system factor, both in biological and synthetic systems. Classically, communication of information over a longer distance is expensive in power; a good summary for these approaches is written elsewhere [4]. The capacitance for a line is a function of the distance of the connection, as well as making connections from one package to another or making connections between boards or other approaches.

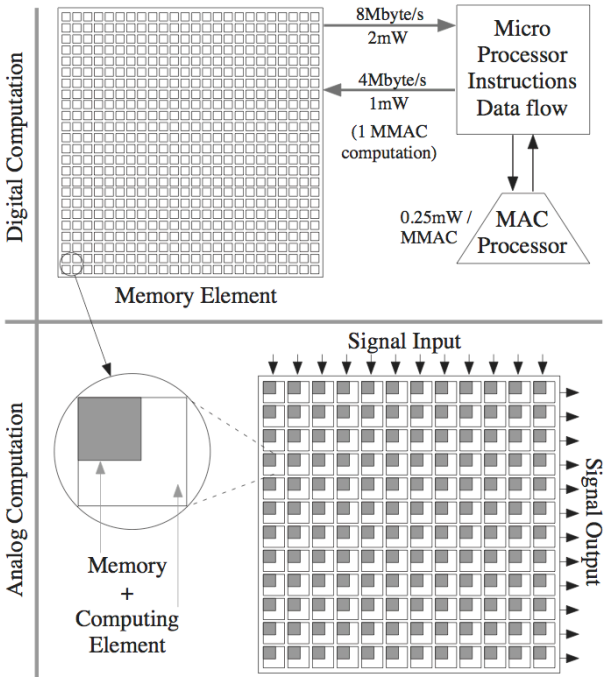

Fig. 1. Diagram showing typical computation models for digital and analog approaches. For a typical digital computation, we must access the data (as well as instructions), communicate it to the processor, perform the computation, and communicate the results back to the memory. When this memory is an off- chip device, the resulting power consumed for communication is much higher than an efficient processor. The analog approach directly computes through the memory, and therefore minimizes the resulting issues and complexity due to communication. One could use digital based computation and memory to achieve some advantages, limited by the computational efficiency limits for digital techniques.

Figure 1 shows an example where the computation power to access 1MMAC of data from a nearby memory block, requiring two 2Mbyte, 32bit input data, and 1Mbyte, 32bit output data, results in 3.1mW (Vdd = 2.5V) of power, even though one might find a DSP chip computing at 4MMAC(/s)/mW power efficiency [5], close to the power / energy efficiency wall [6]. A memory chip or data source further away requires even higher level of power. As another example, using a memory element one chip away for remapping neuron addresses, which is usually a first step to storing synaptic weights in off-chip memory, requires sending an 8bit address off the chip and an 8bit address back on the chip. Just this power alone requires 0.5nJ per remapping in the best case; at 10^{12} events / s, we require 500W for this simple computation. Such an expensive computation must be used in particular targeted areas.

Digital Computation of Events

Figure 2 shows that, where possible, we want to have as much communication locally on a single IC for low-power operation, since that decreases the total amount of capacitance needed to be charged and discharged (i.e., 1 pF for long distance connection on chip), as well as allows for a (lower) range of V_{dd} could be supplied as well as a range of possible communication schemes. Integrating memory and computation, as in biological systems also keeps communication power manageable.

Figure 2a shows a few representative levels for communication of events, typical boundary locations for typical communication. Where possible, we want to have as much communication locally on a single IC for low-power operation, since that decreases the total amount of capacitance needed to be charged and discharged (i.e. 1pF for long distance connection on chip), as well as allows for a (lower) range of V_{dd} could be supplied as well as a range of possible communication schemes. Further, the tighter integration between memory elements and computation further decreases communication power. The types of approaches at a local level needed to optimize the use of memory in the routing architecture. Dendritic structures bring more of the information refinement to the axon outputs.

Almost all systems require communication between multiple chips. When communicating events with a neighbor chip (e.g. 1 chip right next to the transmitting IC), the minimum capacitance is typically set by 10pF by specification (due to packaging, bonding, etc.), as well as off chip communication tends to be at larger V these calculations), resulting in a higher energy computation. Such an approach results in 31.3pJ per bit (or 31.3µW/(Mbit/s)) independent of the communication scheme. Such event communication schemes could transmit an event in only a single bit on the resulting line. Further, the introduction of 3D silicon processing (die stacking, multiple grown layers, etc) has introduced technologies that can reduce the effective off chip capacitance by an order of magnitude, and therefore, such

approaches should be utilized where available in a particular technology for multichip approaches.

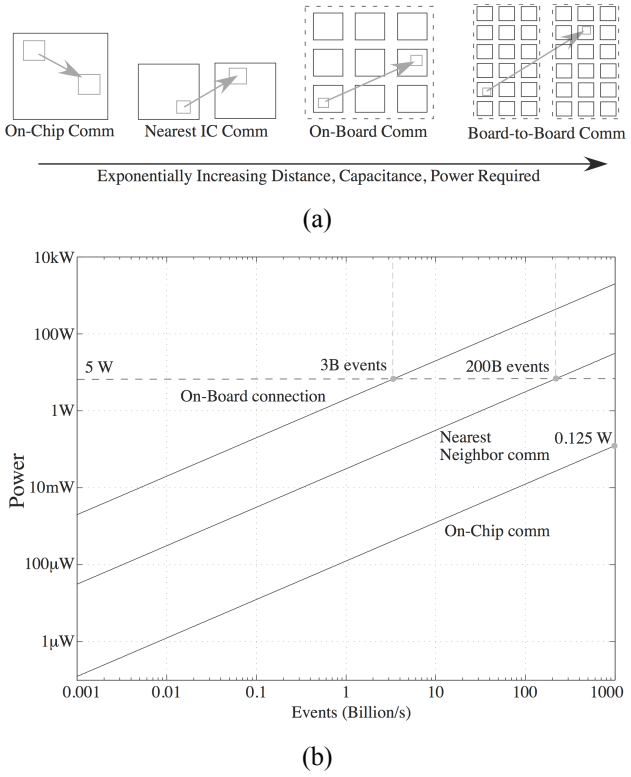

Figure 2: Modeling of power required for transmitting an event. (a) We consider computation between devices on a single IC, between neighboring ICs, on a single board, and distances beyond a single board (i.e. between two boards). Each of these steps requires considerably more power for communicating the resulting event; the more local the communication, the more power efficient the resulting computation. (b) Communication power versus number of events (Gbit) communicated. We consider the three cases of transmitting a bit on a chip (average C_L = 1pF, V_{dd} = 0.5V), transmitting a bit to a neighboring chip (average C_L = 10pF, V_{dd} = 2.5V), and transmitting an event address of 8 bits on a board (average C_L = 80pF, V_{dd} = 2.5V). Each case requires 0.12pJ, 31.3pJ, and 2nJ energy communication per bit, respectively. We would expect even more power consumption for longer distance communication (i.e. between boards), because of the larger capacitance for these approaches. On board requires address communication, because when transmitting sparse events encoding the address gives an optimal solution.

When we communicate over distances longer than nearest neighbor chips, we typically employ an Address Event communication scheme (i.e. AER), which requires sending the location of a particular spike between chips. At least, this requires an address for the particular line, as well as the particular chip we are considering; on a single board, an 8bit address would be a lower limit for such approaches. In such an approach, a communication of an event would travel multiple minimum chip distances (i.e. 8 is a lower bound for an average number), resulting in roughly 2nJ per operation. As we go to longer distances, and particularly when we go to different boards, we see a significant increase in capacitance and addressing as well as routing infrastructure; the goal is to minimize the number of such long distance events that need to be communicated, while preserving the capability.

Figure 2b shows a graph of the power required for communicating a number of events for these different schemes. When trying to reach biological efficiencies for communication, we have significant limits even communicating single events between neighboring ICs, not to mention longer distance communication. 10^{12} events per second results in 30W of power consumption (1 Tbit/s). The result requires most of the computation to be local; fortunately, neurobiological systems use a similar approach in the fact that over 90% of neurons in cortex project locally to nearby neurons (i.e. nearest 1000 pyramidal cells).

For example, if the off chip (not nearest neighbor communication) to is budgeted for 1W of power, then only 0.05% of events can use this communication channel. Further, if we budget 1W for off-board events, then with the additional capacitance and bits for selection needed, one would see 64 times more capacitance, resulting in 0.001% events communicating off board. As additional technology becomes available, such as multiple die stacking in a given package or three-dimensional circuit fabrication, the resulting capacitance for communication will decrease, improving some of these numbers, but the containing concepts will still be the same. We expect similar type issues in neurobiological systems; even though the brain can communicate over long distances by many wires, the resulting energy to do so would be prohibitive in its current energy budget. Such constraints keep the communication overhead for the system manageable, and therefore the communication structure never becomes too large a burden for the system scaling to large sizes.

The low spike rate has a similar effect for synthetic systems as it does in biological systems; increasing spike rate by a factor of 100, typically necessary for implementations using rate encoded approaches, increases power by at least a factor of 100, significantly limiting where such systems can be used. Of course, most rate encoding approaches simplify neuron elements to elementary sigma-delta converters, eliminating most of the computational possibilities.

Example of Communication in a small Network

Figure 3 shows the tradeoffs between these systems, as well as simple comparisons between a small network of simple neurons and synapses. Using external memory as the primary approach for programmability and configurability, as is the typical use of Address-Event Representation (AER) communication schemes, comes at

a huge cost that makes scaling to large systems impractical.

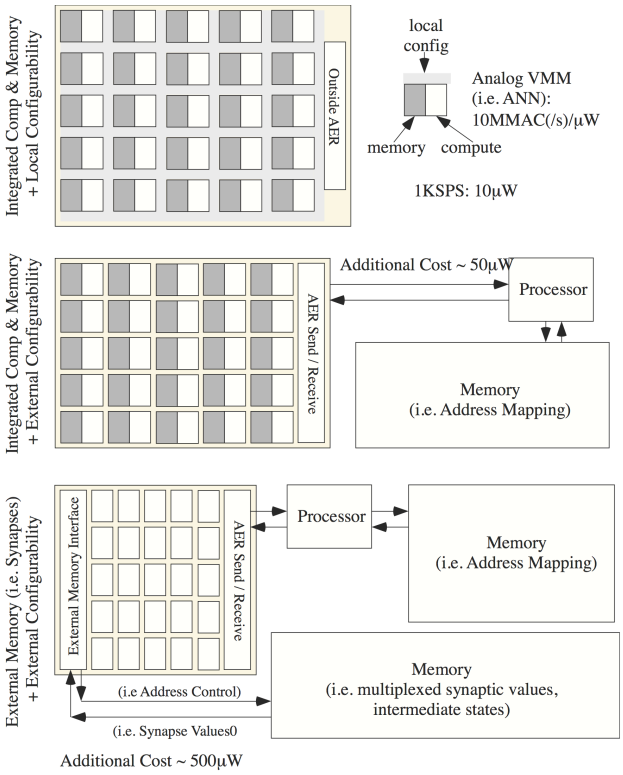

Figure 3: Illustration of the costs of external communication for configurability and storage. Where possible, we want data-flow operations where memory and computation are co-located with local routing / configurability. Moving configurability is moved off of the processing die substantially increases computational cost because of the power and complexity requirements for moving the data to an external processor / memory, even if next to the IC. Moving memory away from Processing, say for multiplexing Synaptic values, further increases the resulting power and complexity cost, even if the original device gets simpler and smaller. These schemes include rate-encoded approaches encoding synapse values because of the increased event rate. We include values for a small network of 1000 neurons with 100 synapses operating with a 1KSPS operating speed assuming a typical ANN (i.e. Vector-Matrix Multiplication) neuron structure.

The advantages of AER communication, which include enabling long-range, sparse interconnections, comes with the added cost of digital communication, costs that are very small for sparse, infrequent events, and that depend on the distance required for communication (on-chip, off-chip, off-board). Adding the additional cost of FPGA or other high performance digital processing only further weakens the applicability of these approaches going forward. One sees exactly the same issue when using multiplexing of a memory with an analog system, whether to load synaptic weights in an external memory. This result shows the heavy energy cost of computation and memory that are not co-located; although this approach might have advantages in initial system building, it requires communication across sizable capacitance, and therefore requiring more power, as well as system complexity.

Many neuromorphic systems claim to be power efficient, and compared to typical digital off-the-shelf approaches, these claims are often right. In each of these approaches, the IC power efficiency is between the digital and analog SP techniques, with much lower system power efficiency due to the high-level for communication overhead (including FPGAs for routing). Many techniques start with a power efficient neuromorphic sensor, such as the DVS imager [7], which compares well to commercial cameras, making it a favorite sensor interface for many neuromorphic platforms. Unfortunately, neuromorphic techniques have not often improved past the analog SP efficiency; often the approaches, including event- based approaches, reduce down to Vector-Matrix Multiply operations, as sometimes explicitly said by the authors [8]. Any practical neural implementation must make sure that the resulting infrastructure does not overwhelm the efficient computation, considering system communication of events, communication to outside processors, and other multiplexing structures. These facts leave us with a small list of potential neuromorphic computational models currently used; the authors believe more efficient algorithms will be discovered / invented over the coming years.

References

[1] J. Hasler and B. Marr, “Finding a roadmap to achieve large neuromorphic hardware systems,” *Frontiers in Neuroscience*, vol. 7, no. 118, 2013.

[2] R.J. Douglas and K.A.C. Martin, “Neuronal Circuits of the Neocortex,” *Annual Rev. Neuroscience*, vol. 27, 2004, pp. 419-451.

[3] A.L. Hodgkin, and A.F. Huxley, and B. Katz, “Measurements of current-voltage relations in the membrane of the giant axon of Loligo,” Journal of Physiology, vol. 116, no. 4, 1952, pp. 424-448.

[4] E. Culurciello, and A.G. Andreou, “Capacitive inter-chip data and power transfer for 3-D VLSI”, *IEEE Transactions on Circuits and Systems II,* 2006.

[5] http://www.ti.com/product/tms320vc5416

[6] H. B. Marr, B. Degnan, P. Hasler, and D. Anderson, ``Scaling Energy Per Operation via an Asynchronous Pipeline," *IEEE Trans. on VLSI*, Vol. 21, No. 1, January 2013, pp. 147-151.

[7] P. Lichtsteiner, C. Posch and T. Delbruck, “A 128x128 120dB 15us Latency Asynchronous Temporal Contrast Vision Sensor,” IEEE Journal of Solid State Circuits, vol. 43, no. 2, 2008, pp. 566-576.

[8] R. Serrano-Gotarredona, et. al, “CAVIAR: A 45k Neuron, 5M Synapse, 12G Connects/s AER Hardware Sensory–Processing- Learning–Actuating System for High-Speed Visual Object Recognition and Tracking,” IEEE Transactions on Neural Networks, vol. 20, no. 9, Sept. 2009.