Massive Parallelism in Neurocomputation

Bruce J. MacLennan

Computer Science Department
University of Tennessee, Knoxville
maclennan@cs.utk.edu

May 23, 1994

Abstract

Neurocomputing is the attempt to apply the principles of information processing that occur in the brain, as they are currently understood.

However, the characteristic of neurocomputing are not very compatible with contemporary scientific supercomputers.

We argue that full exploitation of neurocomputation will require the design of massively parallel analog computers with a million to a billion processing elements and a billion to a trillion interconnections.

The unique characteristics of neurocomputing make such computers feasible.
1 Introduction

Many applications require the flexible processing of large amounts of ambiguous, incomplete, or redundant information, including images, speech and natural language.

Recent advances have shown that many of these problems can be effectively solved by artificial neural networks, which are simplified models of information processing in the brain.

Massively parallel scientific computers might seem a good choice for implementing artificial neural networks (and they often are), but we will argue that conventional computing, as exemplified by scientific computing, contrasts with neurocomputing in several ways:

<table>
<thead>
<tr>
<th></th>
<th>Conventional</th>
<th>Neural</th>
</tr>
</thead>
<tbody>
<tr>
<td>computation</td>
<td>narrow but deep</td>
<td>wide but shallow</td>
</tr>
<tr>
<td>state space</td>
<td>high precision digital</td>
<td>low precision analog</td>
</tr>
<tr>
<td>processors</td>
<td>complex and fast</td>
<td>simple and slow</td>
</tr>
<tr>
<td>representations</td>
<td>localized</td>
<td>distributed</td>
</tr>
</tbody>
</table>

Each kind of computing has its characteristic strengths and weaknesses [17].

Distinct advantages of conventional information processing are:

- definiteness
- reliability
- reproducibility
Distinct advantages of neural information processing are:

- flexibility
- robustness, adaptiveness
- responsiveness

In these posters we compare and contrast massive parallelism in scientific computing and neurocomputing, and argue for the implementation of neurocomputers with very large numbers of simple, analog processors.

Definition:

*Neurocomputing* is the attempt to apply the principles of information processing that occur in the brain, as they are currently understood.

The following terms are approximately synonymous: neural networks, connectionism, neurocomputing.

Neurocomputing is an unconventional style of computing, but mostly simulated on conventional computers.
Why imitate the brain?

At present there are still many important processes that brains do much better than conventional computers:

- flexibility in the presence of ambiguous, incomplete, noisy, novel information
- context sensitivity
- flexible pattern recognition
- flexible control

To better understand the differences between neurocomputing and conventional computing, we’ll first review some characteristics of information processing in the brain.
2 Information Processing in the Brain

Neural information processing is very different from conventional computation.

For example, the “100 Step Rule” [5] says that if we divide the time it takes a person to perform a simple perceptual task, such as recognizing a familiar face ($\sim 1$ sec.), by the response time of a neuron ($\sim 10$ msec.), we find that there can be at most about 100 sequential steps between the stimulus and the response.

A conventional computer cannot do much in 100 steps, certainly not something as complicated as recognizing a face.

One reason that brains are able to accomplish this is that they use massive parallelism, but on a scale well beyond the capabilities of current supercomputers.

For example, even in the first stage of visual processing we have approximately 100 million receptor cells driving one million ganglion cells, each of which is an independent, but slow-speed, low-precision analog processor.

But the story does not end with the large number of neurons involved in information processing.

Although the simplest model of neurons treats the dendrites like simple input wires, there is considerable evidence that the dendritic net may be the site of most of the brain’s information processing [22, 15].

Furthermore, simple neural net models treat neurons as though they
change state discretely, whereas it is possible that interference of continuously varying electrochemical waves is essential to the brain’s pattern recognition capabilities [21].

We have shown elsewhere that sophisticated spatiotemporal information processing that can be accomplished by a single neuron or by small groups of neurons [13, 15, 16, 18].

Next we’ll consider in more detail the characteristics of information processing in the brain.

Brain Structure:

- Topologically, the brain is a two-dimensional sheet (7 layers)
- Local connections in an area are within the sheet
- The topology of the sheet is significant (i.e., information that is stored together is close in abstract space)
- There are nonlocal “projections” from one area to another
- These projections are “topology-preserving” (e.g. mapping retinal images to retinal images)

Temporal Processing

The temporal relations of neural events are significant to information processing (Fig. 1).
Figure 1: Temporal summation of postsynaptic currents resulting from presynaptic action potentials.

Postsynaptic current \((g)\) measured relative to maximum conductance resulting from single AP, time \((t)\) relative to membrane time constant \(\tau\) (8.2 msec. is typical).

Four action potentials arrive at intervals of size \(\tau\) (i.e., at \(t = 0, 1, 2, 3\)), i.e., at 122 Hz rate.

Notice that although there are four APs, the maximum total conductance is \(g \sim 2.5\), not \(g = 4\) as would occur with simultaneous APs.

Based on idealized model: \(g(t) = \alpha^2te^{-\alpha t}\), with \(\tau = 1/\alpha\).
Local Circuits in the Dendritic Net

There is increasing evidence that the basic computational units of the brain are not neurons, but synapses.

For example, Fig. 2 shows a typical arrangement of three synapses that implement a negative feedback loop (essentially an operational amplifier) with an external gain control.

Computation in the Neuropil

Figure 3 is a depiction of the neuropil, the dense “feltwork” of dendrites that constitutes much of the neural cortex.

The neuropil provides a rich medium for the spatiotemporal interaction of analog signals.

Recurrent Filter Model of Neuropil

Figure 4 shows a model of neural information processing that takes account of the spatiotemporal processing capabilities of the neuropil.

Sections of neuropil are viewed as recurrent filters, with each synapse representing one or more variables.

(Many neurons in the cortex have 1000 to 10 000 synapses each.)
Figure 2: Local Circuits in the Dendritic Net.

The lower diagram shows a typical neuron in schematic form.

The “hairy” projections on the dendrites represent dendritic spines.

Many neurons have thousands of dendritic spines, which may be the most basic computational elements of the nervous system.

The circled area is enlarged in the upper diagram.

An axon coming in from the left makes an excitatory synapse on the dendritic spine on the right.

The dendritic spine in turn make an inhibitory synapse back onto the axon terminal, thus setting up a negative feedback loop.

The synapse from the axon entering at the top modulates the action of the first axon.
Figure 3: Depiction of cross-section through neuropil (rat cerebral cortex), magnified $\sim 3500$ times.

Darker areas represent axons.

Lighter areas represent dendrites.

\begin{center}
\textit{Neural Computation is Slow}
\end{center}

Nerve impulses are approximately 1 – 10 msec. long, so the maximum rate is $\sim 100 – 1000$ Hz.

The delay of a chemical synapse is $\sim 0.5$ msec.

(An electrical synapse is essentially delayless, but they are rare in mammalian nervous systems.)
Figure 4: “Resonance Model” of neural information processing.

The neuropil is viewed as a very large number of electrically and chemically coupled oscillators.

Incoming impulses drive the dendritic net, activating various resonances.

Currents are integrated in the dendritic trunks, which act like probes embedded in the neuropil.

Currents reaching the cell body and axon hillock may trigger impulses that are transmitted to other areas.
Representation: “Analog” vs. “Digital”

For the most part, the brain is an analog computing device.

Information is represented on the axons by the impulse rate, which is continuously variable.

In dendritic interactions, information is represented by graded potentials and other continuously variable quantities (such as ion currents and neurotransmitter concentrations).

Stored information (memory) is represented by synaptic efficacy, the ease with which signal cross the synaptic cleft.

Synaptic efficacy is continuously variable, since it is controlled by the physical dimensions of the synapse and the concentration of receptor sites.

Precision in the Nervous System

The nervous system is made from low precision computing devices.
Axonal Signaling

The precision of axonal signaling is limited by the Gabor Uncertainty Principle. It implies:

With 1 kHz bandwidth, $N$ msec. are required to distinguish $N$ values.\textsuperscript{1}

Hence there is at most 1 – 2 digits precision in axonal signals.

Chemical Synapses

The precision of chemical synapses is limited by the fact that releases of neurotransmitter are quantal.

An action potential releases 100–200 quanta.

Hence, there are at most about 2 digits of precision.

\textsuperscript{1} In fact, 100 Hz. is probably a better estimate of the bandwidth.


*Electrical Synapses*

Electrical synapses are not well understood, and seem to be rare in mammalian brains.

The quantal change in conductance is about 5 times that of a chemical synapse.

This suggests that electrical synapses are even lower precision (but quicker) than chemical synapses.

*Conclusion:*

1% – 10% is “close enough for brain work”
Robustness Principle

As a consequence of low precision computation and the prevalence of noise, we know:

- Neural computation cannot depend on an absolute mathematical property.
- Example: exact orthogonality can’t be necessary.
- Example: finite support can’t be necessary.
- Example: completeness can’t be necessary.
- Example: exact linearity can’t be necessary.

That is, many mathematically important issues are irrelevant to neurocomputation.

Massive Parallelism

It is important to appreciate the degree of parallelism used by the brain; see Fig. 5 and Table 1.

In each square centimeter there are approximately 100 billion active computing elements, which drive about 15 million output transducers.
Figure 5: Neurons and synapses in 1 cm² of cortex.

To help visualize the density of neurons and synapses in the brain, the above square represents a 1 cm² section of mammalian cortex, which contains approximately $15 \times 10^6$ neurons and $10^{11}$ synapses.

These numbers are nearly constant throughout the brain, no matter what the thickness of the cortex (1–3 mm is typical).

Primate visual cortex is an exception in that it has over twice the density of neurons as other cortex.

---

*Scale of Dendritic Computation:*

- Significant computing takes place in dendrites
- There are microcircuits in individual dendritic spines
- A synapse is about 0.5 – 2.0 μm across
- There are about 300 million ($300 \times 10^6$) synapses per mm³
Massive Parallelism in the Human Brain:

- One trillion ($\sim 10^{12}$) neurons
- One quadrillion ($\sim 10^{15}$) synapses
- Eye: $\sim$ 100 million receptor cells
- and $\sim$ 1 million ganglion cells


Summary: Conventional Computing vs. Neurocomputing

- Conventional Computing
  - narrow but deep
  - high precision digital
  - complex, fast processors
- Neurocomputing
  - wide but shallow
  - low precision analog
  - simple, slow processors
Capabilities of Neural Computation

We will briefly consider the capabilities of neural computation, based on our own model and a model developed by Bartlett Mel of CalTech.

Linear System Analysis

We argue (Appendix A) that in many cases the neuropil approximates a linear system with a very large number of variables.

Dendritic Network in Terms of Interconnection Matrices

To a first approximation, a dendritic net can be treated as a linear system defined in terms of four matrices of synaptic efficacies (connection weights); see Fig. 6.
Figure 6: Dendritic network in terms of interconnection matrices.

The vector $\mathbf{x}$ represents the axonal input signals, and the vector $\mathbf{y}$ represents the axonal output signals.

The vector $\dot{\mathbf{u}}$ represents the differential feedback in the dendritic net.

Interconnection strengths between inputs, outputs and state variables are represented by the matrices $\mathbf{D}$, $\mathbf{E}$, $\mathbf{F}$ and $\mathbf{G}$.

Dynamics are defined by state equation $\dot{\mathbf{u}}(t) = \mathbf{D}\mathbf{x}(t) + \mathbf{F}\mathbf{u}(t)$, and output equation $\mathbf{y}(t) = \mathbf{E}\mathbf{x}(t) + \mathbf{G}\mathbf{u}(t)$. 
Potential Number of Resonances

- The Laplace-transformed transition matrix will have poles (infinities) at the eigenvalues of the feedback matrix $F$.
- These are its resonances.
- The feedback matrix may have as many eigenvalues as the number of state variables.
- Number of resonances $= \mathcal{O}(10000)$.
- This $suggests$ the pattern recognition capacity of dendritic net.

Pattern Associating Capacity of a $5 \times 5$ mm Slab of Neocortex

- Assumptions:
  - Patterns are 10 000-variable vectors
  - 10 000 minicolumns, 50 neurons each
  - Each computes one output variable
- Allowing for
  - multiple eigenvalues
  - redundancy
- $\mathcal{O}(100 000)$ patterns can be associated
Bartlett Mel’s Analysis

Based on a combination of theoretical analysis and simulations based on biological data, Bartlett Mel (Caltech) has estimated [20]:

- $5 \times 5$ mm slab of neocortex can associate $O(100000)$ pattern pairs
- patterns are $10\,000$ bits
- of which $1000$ are $1$
- $5 \times 5$ mm slab contains $5 \times 10^5$ neurons and $5 \times 10^9$ synapses

This agrees with our estimate (though the model is quite different).
Table 1: Approximate Numbers of Neurons\textsuperscript{2}

\begin{center}
\begin{tabular}{ll}
human & $\sim 30 \times 10^9$ neurons \\
rat & $\sim 65 \times 10^6$ neurons \\
1 sq. mm. cortex & $\sim 150 \times 10^3$ neurons \\
& $\sim 10^9$ synapses
\end{tabular}
\end{center}

3 Characteristics of Neurocomputing

We have argued that achieving the promise of neurocomputing will require a new kind of computer, with principles of operation more like those of the brain.

Thus, instead of a comparatively small number of fast, high precision processors, future neurocomputers should make use of a very large number of possibly slow, low precision processors.

On the other hand, even an animal as simple as a bee may have a million neurons in its brain, and the human brain is estimated to comprise thirty billion neurons (Table 1).

It is not unreasonable to suppose that if we want our neurocomputers to behave at least as intelligently as bees, then they will have to be capable of implementing neural networks of comparable size.

\textsuperscript{2}Sources are Changeux [2, pp. 50–51] and Shepherd [23, p. 392].

An area of mammalian cortex of size $25\mu \times 30\mu$ contains approximately 110 neurons, regardless of the thickness of the cortex (though in primates the visual cortex has 2 to $2\frac{1}{2}$ times this number).

We are assuming 6000–7000 synapses per neuron.

The DARPA Neural Network Study [4] suggests that a bee’s brain has approximately a million neurons, but this seems inconsistent with the data we have, and the DARPA book cites no source.
Thus, the hardware goal for full exploitation of neurocomputing should be to implement between a million and a billion neurons with at least the speed and precision of biological neurons.

We put this conclusion in the form of a strong claim:

If an implementation technology will not scale up to a million to a billion processors with a billion to a trillion connections, then it is not on the trajectory to true neurocomputing.
Current Implementations of Neurocomputing

- Software emulation
- Supercomputers
- Accelerator boards
- Analog VLSI

These are useful, but fall far short of permitting brain-scale parallelism.

Computational Capabilities of the Brain

- 1 cm² of cortex has $10^{11}$ synapses
- Realtime simulation requires updating at least 1/msec.
- Therefore 100 teraflops ($100 \times 10^{12}$ flops) are required to equal power of each square centimeter of cortex
- Current supercomputers: $\sim 10$ gigaflops
- Current supercomputers fall short by approximately four orders of magnitude
- Furthermore, we would like this computational power in a package of comparable size, i.e., 1 cm²
4 Neurocomputing Hardware

We will argue that massively parallel analog neurocomputers are feasible because:

1. Low precision analog hardware is fundamentally simpler than digital hardware.
   For example, Carver Mead has shown that analog VLSI technology can implement a signed analog adder with 4 transistors, a signed multiplier with 9 transistors, and a square-root with 3 transistors [19, Ch. 6].
   A digital computer requires 2 or more transistors for each flip-flop.

2. The individual processors required for neurocomputing are very simple: the most common model uses saturating linear combinations.

3. Since neurocomputation is “wide” rather than “deep,” analog neurocomputation can be implemented by much slower processes (e.g. combinations of electrical, optical and chemical processes) than those needed for digital, scientific computation.

4. Neural information processing represents information in a distributed way, which takes advantage of redundancy to achieve reliability.
   This, especially when combined with the principles of “field computation” [7, 8, 9, 10, 11, 24] allow the use of irregular processor arrays, which are amenable to less precise manufacturing techniques than those required for VLSI.
   As a consequence, much higher degrees of parallelism will be achievable.
All of the preceding factors will simplify implementation of massively parallel, low precision analog neurocomputers.

In summary, the following characteristics of neurocomputing make the implementation of brain-scale parallelism feasible:

- Slow speed components
- Low precision computation
- Tolerance of imprecise layout
- Error and failure tolerance
- More “Natural”:
  - Nature is parallel
  - Laws of nature are “analog” (continuous)
  - High precision not required

Conclusions

By taking advantage of the unique characteristics of neurocomputing, we may be able to design a neurocomputer technology that is much more efficient — for neurocomputing — than conventional supercomputer technology.
A In the Neuropil Linear?

_synapses are approximately linear:_

- A chemical synapse can be treated as a voltage-controlled voltage-source.
- At low neurotransmitter concentrations postsynaptic conductance depends linearly on presynaptic depolarization.
- At high neurotransmitter concentrations receptors saturate and dependence becomes sublinear.
- Overall: transfer is sigmoidal with a large linear region (Figs. 7, 8).
Figure 7: Peak postsynaptic conductance as a function of postsynaptic potential.

Fraction $\theta$ of the maximum conductance plotted against $V$, potential relative to reversal potential: $\theta = 1/(1 + e^{-kV})$, where $k = 3.9$ mV (sodium conductance [19, pp. 53–57]).
Figure 8: Spike rate versus input.

(redrawn from [1])

(a) Reciprocal spike interval versus depolarization current for a Class I crab axon (Chapman, 1966).

(b) Discharge rate versus test spot luminance for retinal on-center ganglion cell, for differing adaptation luminances (Creutzfeldt, 1972).
References


