# von Neumann's missing "Second Draft": what it should contain

János Végh
*Kalimános BT*
Debrecen, Hungary
Vegh.Janos@gmail.com ORCID: 0000-0002-3247-7810

***Abstract*—Computing science is based on a computing paradigm that is not valid anymore for today's technological conditions. The reason is that the transmission time even inside the processor chip, but especially between the system's components, is not negligible anymore. The paper introduces a quantitative measure for dispersion, which is vital for computing performance and energy consumption, and demonstrates how its value increased with the changing technology. The temporal behavior (including the dispersion of the commonly used synchronization clock time) of computing components has a critical impact on the system's performance at all levels, as demonstrated from gate-level operation to supercomputing. The same effect limits the utility of the researched new materials/effects if the related transfer time cannot be proportionally mitigated. Von Neumann's model is perfect, but now it is used outside of its range of validity. The correct procedure to consider the transfer time for the present technological background is also derived.**

***Index Terms*—modern computing paradigm, performance limitation, efficiency, parallelized computing, supercomputing, high-performance computing, von Neumann architecture**

## I. Introduction

Today's computing is commonly thought to be based on the famous "First Draft" [1] by von Neumann[1]. In that report, he discussed the principles and the technical implementation of his paradigm in parallel with the brain's neuronal operation, so his model was brain-inspired. He emphasized the computing process's timing relations and analyzed the role of delays in the computing chain. He also called attention to the fact that *the transfer time is an integral part of the computation process*: whether it is biological, digital, or analog, processing cannot even begin before its input operands are delivered to the place of operation, and vice versa, the output operand cannot be delivered until the computing process terminates. *The transfer time and the operation time mutually block each other in a computing chain*.

Given that *he considered the intended vacuum tube implementation*, to simplify the model, *he proposed to neglect the transfer time*, aside from the processing time. In other words, his approximation assumes instant interaction between his "computing organs". However, he explicitly warned that it would be *unsound* to use that classic paradigm, neglecting the conduction time, to describe neuronal operation, given that the conduction time is longer than the synaptic time.

Another condition that –with his words– *vitiates* his paradigm, is if the processor is "too fast". If we consider a 300 $m^2$ sized computer room and the 3000 vacuum tubes estimated, von Neumann considered a distance between vacuum tubes about 30 cm as a critical value. At this distance, the transfer time is about three orders of magnitude lower than the processing time (between the "steps" he mentioned). These limitations are why von Neumann justified the procedure *for vacuum tube technology only*. He noted that using a hundred-fold higher frequency, even with vacuum tubes, *vitiates the neglection* he proposed. At such a frequency, the transfer time approaches (or even exceeds) the order of magnitude of the processing time, so neglecting the transfer time cannot anymore be justified: the *apparent processing time* (the clock time between consecutive computing operations) differs from the *physical processing time* by dozens of percent.

## II. The history of temporal characteristics

Today, the size of a processor chip is in the three cm range, and correspondingly, the transfer time can reach the 1 ns range. The processing time (in the sense as von Neumann used the term) is well below the range of 1 ns, so it is worth to check if today's computing can be based on the classic paradigm. "*Building this new hardware [neuromorphic computing] necessitates reinventing electronics*"; "*more physics and materials needed*" [2]. It is not sufficient, however. At least ***some physics is needed even in the computing paradigm***, to benefit from the researched new effects and materials. It was noticed that "*computers have undergone tremendous improvements in performance over the last 60 years, but those improvements have significantly slowed down over the last decade, owing to fundamental limits in the underlying computing primitives*" [3]. However, the real reason is at one level deeper: *we are using the otherwise correct computing model under technological conditions where the neglections used in the classic paradigm are not valid any more*.

Given that von Neumann derived his neglections *apart from processing time*, we derive some quantitative merits, as discussed below. Fig. 1 shows their dependence on the year of fabrication of the processor. The technical data are taken from

Project no. 136496 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the K funding scheme.

[1]This publication is commonly referred to as "von Neumann architecture", although von Neumann in the first sentence makes clear that "*The considerations which follow deal with the **structure** of a very high speed automatic digital computing system, and in particular with its **logical control***."

978-1-7281-7624-6/20/$31.00 ©2020 IEEE
DOI 10.1109/CSCI51800.2020.00235

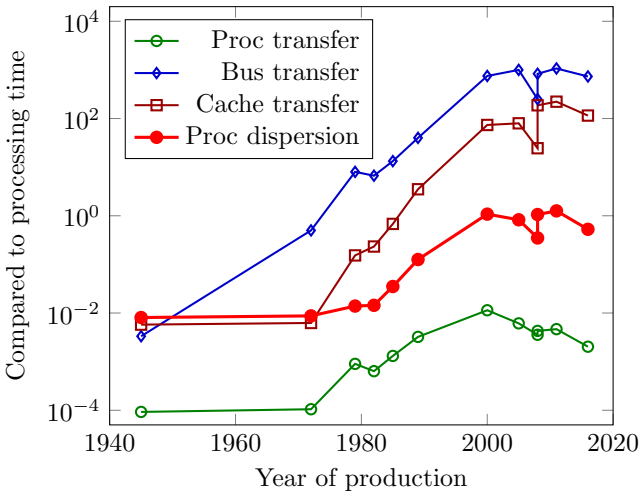


Fig. 1. The history of some relative temporal characteristics of processors, in function of their year of production. Notice how cramming more transistors in a processor changed disadvantageusly their temporal characterisctics.

publicly available data[2] and from [4]. The figures of the merits are results of rough and somewhat arbitrary approximations. However, their consequent use enables us to draw limited validity conclusions without needing proprietary technological data.

We estimate the distance between the processing elements in two different ways. We use the method described above for vacuum tubes to calculate the square root of the processor area divided by the number of transistors. This figure gives a kind of "average distance" of the transistors. We consider it as a minimum distance the signals must travel between transistors[3]. This value is depicted as "Proc transfer" in Fig. 1. The maximum is the distance between the two farthest processing elements on the chip[4].

Given that usually the processing elements and the storage elements are fabricated as separated technological blocks and connected by wires (aka bus), we also estimated a "bus transfer" time. The memory access in this way is extended by the bus transfer time. We assumed that a cache memory could be advantageously positioned at a(n average) distance of half processor size because of this effect. This time is shown as "Cache transfer" time. The cache memories appeared about the end of the 1980s. It became evident that the bus transfer vastly increases the memory transfer time. Using cache memory can enhance systems' performance by order of magnitude (cache data can be calculated for all processors, however).

Given that "The emphasis is on the exclusion of a dispersion" [1], using those technical data, we define a dispersion as the geometric mean of the minimum and maximum "Proc transfer" times, divided by the processing time. In von Neumann's abstraction, the "well-defined dispersionless synaptic delay $\tau$ [processing time]" is used. With our definition, the dispersion of EDVAC is (at or below) 1 %; this is why von Neuman justified his procedure. An interesting parallel is that both EDVAC and Intel 8008 have the same number of processing elements. The relative processor and cache transfer times are in the same order of magnitude. However, notice that the bus transfer time's importance has grown and started to dominate the single-processor performance in personal computers. A decade later, the physical size of the bus necessitated to introduce cache memories. The physical size led to saturation in all relative transfer times.[5] The slight decrease in the relative times in the past years can probably be attributed to the sensitivity of our calculation method to the spread of multi-cores; this suggests to repeat our analysis method with proprietary technological data.

As the "Proc dispersion" diagram line shows, *in the today's technology, the dispersion is near to unity*. This large dispersion means that today's processors' operating regime is more close to the operating regime of our brain than the operating model abstracted in the classic paradigm. However, for our brain, an explicit "spatiotemporal" behavior is considered, and is "unsound" to use the classic paradigm to describe it. That is, we cannot apply the "dispersionless" classic paradigm any more[6]. This high dispersion value is why only a small fragment of the input power can be used for computation; the rest is dissipated (produces heat). We experience it since the dispersion approached unity about two decades ago. *The dispersion of synchronizing the computing operations* vastly increases the cycle time, decreases the utilization of all computing units, and enormously increases the power consumption of computing [6], [7].

In the same section, von Neumann said: "We propose to use the delays $\tau$ as absolute units of time which can be relied upon to synchronize the functions of various parts of the device. The advantages of such an arrangement are immediately plausible". Yes, his statement is valid for the well-defined *dispersionless synaptic delay* $\tau$ he assumed, but not at all for today's processors. The recent activity [2], [3] to consider asynchronous operating modes is motivated by admitting that *the present synchronized working method is disadvantageous in the non-dispersionless world.*

At his time and in the age of vacuum tube technology, von Neumann did not feel the need to discuss what a procedure can justify describing the computing operation in a non-dispersionless case. However, he suggested reconsidering the validity of the neglections he used in his paradigm for any new future technology. To have a firm computing paradigm for the present technologies, *we need to consider the ratio of the transfer time to processing time*; we cannot neglect it anymore. The real question is, the discussion of which is *missing from*

[2] $https://en.wikipedia.org/wiki/Transistor\_count$

[3] Notice that this transfer time also shows a drastic increase with the number of transistors, but alone does not vitiate the classic paradigm

[4] Evidently, introducing clock domains and multi-core processors, shades the picture. However, we cannot provide a more accurate estimation without proprietary technological data

[5] The real cause of the "end of the Moore age", is, that Moore's observation is not valid for the physical bus size

[6] Reaching the plateau of the diagram lines coincides with introducing the "explicitly parallel instruction set computer" [5]: that was the maximum that the classic paradigm enabled.

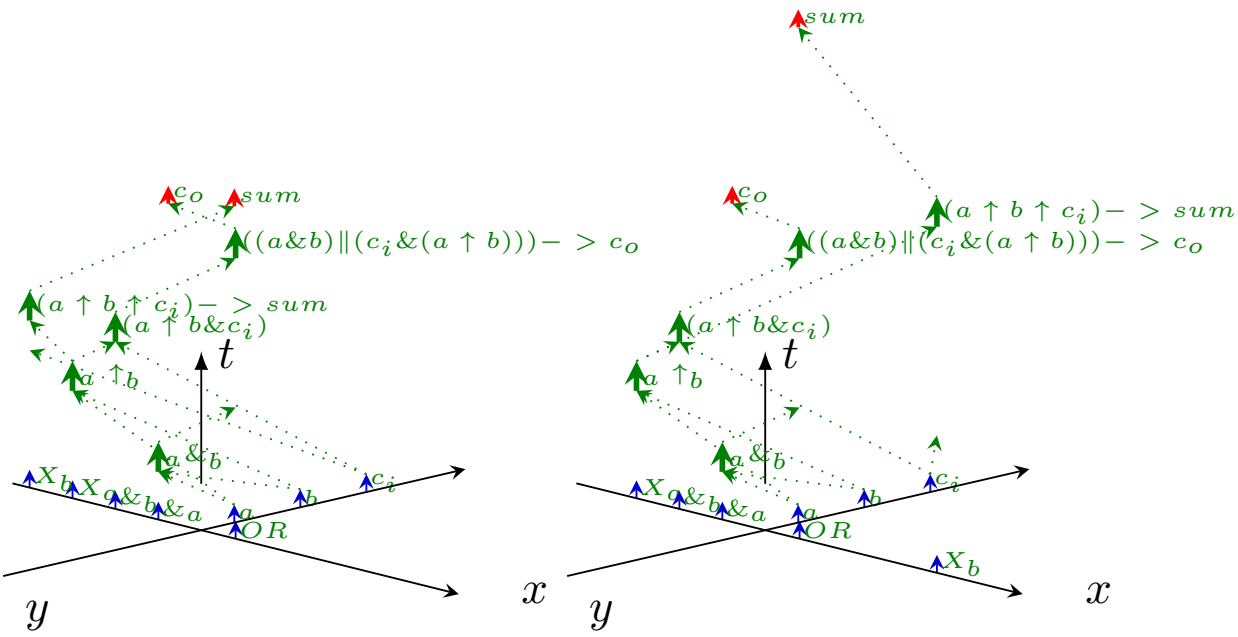


Fig. 2. The temporal diagram of a one-bit adder in the time-space system. The diagram shows the logical equivalent of the SystemC source code of Listing 1, *the time from axis x to the bottom of green arrows* signals "idle waiting" time (undefined gate output). In the left subfigure, the second XOR gate is at (-1,0). In the right subfigure, the second XOR gate is at (+1,0). Notice how changing the position of a gate affects signal timing. Notice also that the lack of vertical green arrows means idle time for the gates.

*the "First draft", what a procedure shall be followed if the transfer time is not negligible*?

## III. Introducing temporal logic into computing

Although von Neumann explicitly mentioned that the propagation speed of electromagnetic waves limits the operating speed of the electronic components, until recently, that effect was not admitted in computing. In contrast, in biology, the "spatio-temporal" behavior was recognized very early. The recent trend is to describe theoretically and model the neuronal operation electronically using the computing paradigm, proposed by von Neumann, which is undoubtedly not valid for today's technology. According to von Neumann, it is doubly $unsound$ if one attempts to mimic neural operation based on a paradigm that is $unsound$ for that goal, on a technological base (other than vacuum tubes) that $vitiates$ the paradigm.

Fortunately, the spatio-temporal behavior suggests a "procedure" that can be followed in the case when the transfer time can even be longer than the processing time. Despite the name "spatio-temporal", biology describes its systems' behavior using separated space and time functions (and, as a consequence, needs ad-hoc suggestions and solutions for different problems). However, it has one common attribute with technical computing: the information transfer speed is limited in both of them. For the first look, it seems to be strange to describe such systems with (the inverse of) the Minkowski transform, given that it became famous in connection with Einstein's theory of special relativity. However, in its original form, only the existence of a limiting speed is assumed.

Listing 1. The essential lines of source code of the one-bit adder implemented in SystemC

```
//We are making a 1-bit addition
aANDb = a.read() & b.read();
aXORb = a.read() ^ b.read();
cinANDaXORb = cin.read() & aXORb;

//Calculate sum and carry out
sum = aXORb ^ cin.read();
cout = aANDb | cinANDaXORb;
```

As discussed in **[?]**, this latter feature enables us to describe the correct behavior of information processing both in science-based technical implementations and biology, for any combination of the transfer time and the processing time. The key idea is to transform the spatial distances between computing components (that can be $Si$ gates, cores, network nodes, biological or artificial neurons) to time (measured with the limiting speed along the signal path).

## IV. Phenomena due to the temporal behavior

Yes, in computing, "more physics ...is needed" [2].

### A. One-bit adder

Describing the temporal operation at gate level is an excellent example, that *the line-by-line compiling (sequential programming, called also Neumann-style programming [10]), formally introduces only logical dependence, but through its*

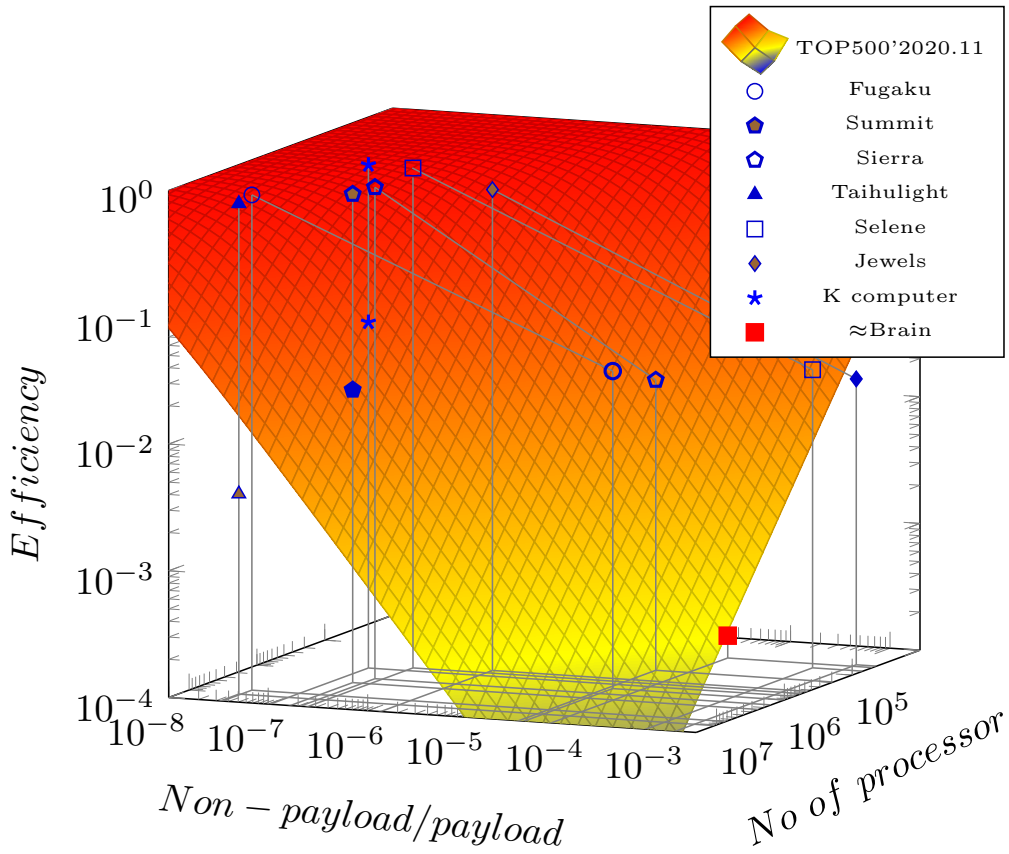


Fig. 3. The surface and the figure marks show at what efficiency the top supercomputers run the 'best workload' benchmark HPL, and the 'real-life load' HPCG. **[8]** The right bottom part displays the expected efficiency **[9]** of running neuromorphic calculations on SPA computers.

*technical implementation it implicitly and inherently introduces a temporal behavior, too*. Fig. 3 in [1] shows a simple adder, in vacuum-tube approach. Listing 1 shows how a one-bit adder is implemented in hardware (HW)-description language SystemC. Fig. 2 shows the corresponding elementary operations, as they happen along the axis $t$ (for a detailed explanation and legend, see **[11]**). Notice that considering their temporal behavior results in longer delays for the adder built from gates used in today's technology, than that for the logical adder assumed by von Neumann. Notice also, what von Neumann called the attention to: *until a gate receives all of its inputs, its output is undefined*. For the drastic consequences of this effect, see the impact of training Artificial Neural Network (ANN)s in **[12]**.

### B. The inherent performance limits of parallelized computing

The temporal behavior of the components leads to a drastic performance loss in the case of parallelized computing. Recall that "*this decay in performance is not a fault of the architecture, but is dictated by the limited parallelism*". [13] Fig. 3 shows the efficiency of some recent supercomputers in function of the number of their processor cores, and their degree of parallelization **[8]**. Notice how the different workload (HPCG) forces to use only a fragment of the available cores.

### C. New effects and materials

Their temporal behavior also limits the usability of new materials/effects, a highly popular idea, especially for neuronal operations [2], [14]. Fig. 4 shows the effect of using a faster cache memory in a computing system. Although the apparent operation gets faster when using a ten-fold quicker (and maybe even more expensive) cache, the reached speedup is not proportional to the speedup in the cache's physical operating speed. For a detailed discussion see **[11]**.

## V. Summary

The present commonly used classic computing paradigm, according to its inventor, is valid for (the timing relations of) vacuum tubes only. It assumes instant interaction between the components of computing systems, which is theoretically wrong and contradicts everyday experience. Concerning their temporal behavior, the computing systems show interesting (but not surprising) parallels with the modern science **[9], [15]**. The phenomena range from the stalled single-processor (used

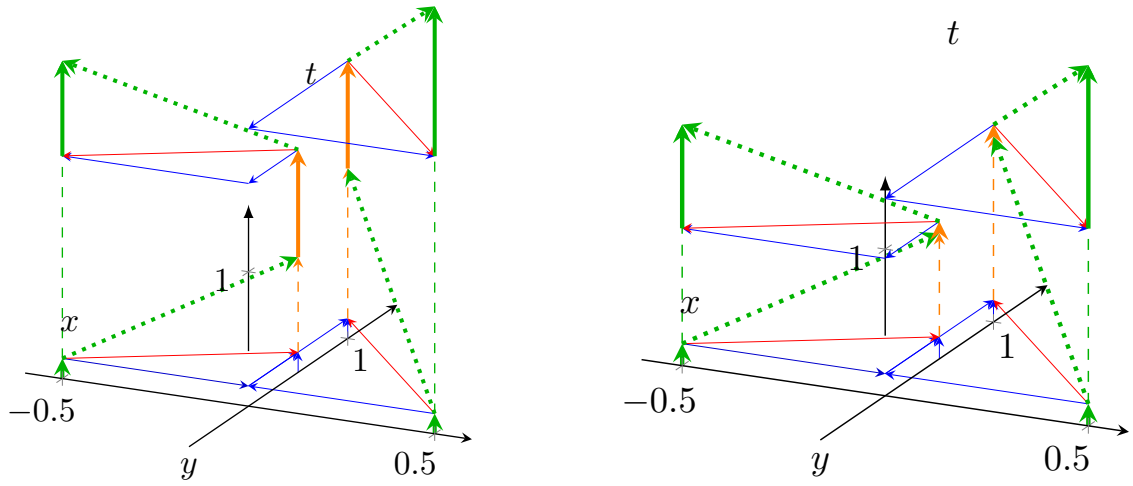


Fig. 4. The performance dependence of an on-chip cache memory, at different cache operating times, in the same topology. The cores at x=-0.5 and x=0.5 positions access the on-chip cache at y=0.5 and y=1.0, respectively. The vertical orange arrows represent the physical cache operating time, and vertical the green arrows the apparent access time. The physical operating speed of the cache memory of the right subfigure is ten times better. Compare the apparent access times to the corresponding physical ones (the time ratio is better only about a factor of two). Notice also that the apparent operating speed is more sensitive to the position rather than to the speed of the cache memory

for heating rather than computing) performance through the supercomputers, unable to exceed their inherent performance rooflines **[8]** and having *payload performance* for real-life task about 1 % of their *nominal performance*, to the stalled scaling of ANNs **[12]** (the latter led to that "AI Core progress has stalled" [16]). Yes, "*building this new hardware necessitates reinventing electronics*" [2].

At this point, we have two options. We can complement the currently existing names for the scientific discipline and its journals so that "Computing Science *for vacuum tubes only*". Furthermore, we can start a new discipline under the name "Modern computing science" (parallel with classic versus modern science) based on temporal logic, describing the computing based on state-of-the-art technology. In this way, we can enter into the "Next Level" [17]. Even *rebooting computing* is not possible without making such a drastic change. The other option is to keep the classic computing science and to admit that the "Game Is Over".

## REFERENCES


[1] J. von Neumann, "First Draft of a Report on the EDVAC," https://web.archive.org/web/20130314123032/ http://qss.stanford.edu/ godfrey/vonNeumann/vnedvac.pdf, 1945.

[2] D. Markovic, A. Mizrahi, D. Querlioz, and J. Grollier, "Physics for neuromorphic computing," *Nature Reviews Physics*, vol. 2, p. 499–510, 2020.

[3] J. D. Kendall and S. Kumar, "The building blocks of a brain-inspired computer," *Appl. Phys. Rev.*, vol. 7, p. 011305, 2020.

[4] J. J. P. Eckert and J. W. Mauchly, "Automatic High-Speed Computing: A Progress Report on the EDVAC," Moore School Library, University of Pennsylvania, Philadephia, Tech. Rep. Report of Work under Contract No. W-670-ORD-4926, Supplement No 4, September 1945.

[5] M. Schlansker and B. Rau, "EPIC: Explicitly Parallel Instruction Computing," *Computer*, vol. 33, no. 2, pp. 37–45, Feb 2000.

[6] R. Waser, Ed., *Advanced Electronics Materials and Novel Devices*, ser. Nanoelectronics and Information Technology. Wiley, 2012.

[7] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and et al., "Dark Silicon and the End of Multicore Scaling," *IEEE Micro*, vol. 32, no. 3, pp. 122–134, 2012.

[8] J. Végh, "Finally, how many efficiencies the supercomputers have?" *The Journal of Supercomputing*, vol. 76, no. 12, pp. 9430–9455, feb 2020. [Online]. Available: http://link.springer.com/article/10.1007/s11227-020-03210-4

[9] J. Végh, "How Amdahl's Law limits performance of large artificial neural networks," *Brain Informatics*, vol. 6, pp. 1–11, 2019. [Online]. Available: https://braininformatics.springeropen.com/ articles/10.1186/s40708-019-0097-2/metrics

[10] J. Backus, "Can Programming Languages Be liberated from the von Neumann Style? A Functional Style and its Algebra of Programs," *Communications of the ACM*, vol. 21, pp. 613—-641, 1978.

[11] J. Végh, "Introducing Temporal Behavior to Computing Science," in *2020 CSCE, Fundamentals of Computing Science*. IEEE, 2020, pp. Accepted FCS2930, in print. [Online]. Available: https://arxiv.org/abs/2006.01128

[12] ——, "Which scaling rule applies to Artificial Neural Networks," in *Computational Intelligence (CSCE) The 22nd Int'l Conf on Artificial Intelligence (ICAI'20)*. IEEE, 2020, pp. Accepted ICA2246, in print; in review in Neural Networks. [Online]. Available: http://arxiv.org/abs/2005.08942

[13] J. P. Singh, J. L. Hennessy, and A. Gupta, "Scaling parallel programs for multiprocessors: Methodology and examples," *Computer*, vol. 26, no. 7, pp. 42–50, Jul. 1993.

[14] "Building brain-inspired computing," *Nature Communications*, vol. 10, no. 12, p. 4838, 2019. [Online]. Available: https://doi.org/10.1038/s41467-019-12521-x

[15] J. Végh and A. Tisan, "The need for modern computing paradigm: Science applied to computing," in *Computational Science and Computational Intelligence CSCI The 25th Int'l Conf on Parallel and Distributed Processing Techniques and Applications*. IEEE, 2019, pp. 1523–1532. [Online]. Available: http://arxiv.org/abs/1908.02651

[16] M. Hutson, "Core progress in AI has stalled in some fields," *Science*, vol. 368, p. 6494/927, 2020.

[17] S. H. Fuller and L. I. Millett, "Computing Performance: Game Over or Next Level?" *Computer*, vol. 44, pp. 31–38, 2011.