Final report on project **K-136496** (formerly K-125547)

During designing the project we have concluded that in designing both single processors and composite computing architectures the so called single-processor approach was used. Symptoms of the case are that the component basis, the operating systems and applications built upon it as well as the knowledge of computing experts are confined within this paradigm. The exponentially growing demand against computerized data processing made more and more obvious that the technological solutions based on that paradigm has reached the limits of its possibilities. The project has started (at the host Miskolc University as project K-125547) originally with the goal to discover – at modeling level – the peculiarities of the operation of cooperating processors and the factors affecting efficiency of their operation. Our goal was to elaborate models for scrutinizing the feasibility of ideas for enhancing their performance and to build a simulator able to demonstrate their effect. In the first year of the project we studied and developed mainly technological solutions in the field. We suggested a multi-port memory solution and a hardware method for eliminating the so called priority inversion. Our activity was acknowledged by an invited talk (the only one from East-Europe) on the Development Summit of the processor manufacturer ARM, in Cambridge, UK. We published our results at leading professional conferences and in journal Parallel Computing.

Unfortunately, the host institution practically made impossible realizing the project (with sending the principal investigator to pension and blocking the access to project resources for nearly two years). In the interest of performing the promised scientific work, the private research company (Kalimános BT, as project 136496) of the Principal Investigator (PI) took over the project and finished it successfully. Because of this forced exchange of host institutions the project had to be modified from several points of view. First of all, accessing the resources of the grant has stalled for nearly two years (and made the planned investments as well as making Open Access publications impossible) and the number of researchers decreased from three to one. The two lost researchers were expected to carry out, among others, the technological (FPGA-modeling) branch. In addition, somewhat later the COVID excluded in-person meetings of researchers, so the focus of the project has been shifted towards the branch of modeling and theoretical research. On the one side, this situation needed single-person research, on the other side the specialty of the PI was modeling technical operation of computing systems and theory of technically implemented computing. The third factor was that a huge interest and research activity (our research contributed significantly) waked up in the field of the so called large-scale computer architecture development, computing and supercomputing, artificial neuromorphic architectures and neural networks, emulating biological neurons by electronic technical and software means. Many technical and biology-mimicking projects failed (in our view: mainly) because of lacking the required theoretical knowledge. Although the co-opted young researcher could not counterbalance the loss of the technical researchers, his biological background and experience was very useful in extending our research direction towards biological computation.

The research started in direction of enhancing computing operation and the solved problems led to another problems in a deeper layer, as the PI comprehended the meaning of the new inventions of the results of our own research. Because of this, for the easier understanding, the rest of the report follows a logical (practically: reversed) time order with respect to the project milestones achieved. Actually, the most valuable theoretical result was discovering the general theory of computing (published in Informatics) and revealing the role of the time behind processes (published in Acta Biotheoretica), which unified the computing paradigm behind technological and biological computing. This discovery led to understanding some mystic issues of technological computing (several conference publications in series “Foundations of Computer Science”), such as the low power efficiency of processors made with modern technology, the low computing efficiency of distributed computing and artificial neural networks (Neural Computing and Applications), the inherent performance limits of supercomputers (J. Supercomputers).

Although von Neumann modeled his computer from the slow, unreliable, parallel - but inexpensive and networked - computing elements of the nature, technology developed fast, reliable, sequential - but expensive and segregated - processors. Despite those evident differences, we started studying biology as researching a field where the theoretically expected effects are more expressed due to the much lower interaction speed. The careful research led to understanding the abstract operation of biological neurons and their network; the neuronal operation, including learning, memory; the notion of neuronal information and its processing. The thorough understanding of these two fields enabled us to make excursions to brain research, artificial intelligence and machine learning, biomorphic architectures.

The origin of our research started from the idea which von Neumann took as a model for his computing paradigm: the human brain. He developed his *computing model* by abstracting the operation of the brain, but, for his practical goal: constructing an abstract interface between (as we call today) hardware and software, he introduced omissions to construct his famous simplified *computing paradigm*. His model was clearly *biology inspired*, but his paradigm was not *biology mimicking:* he introduced the omission that the data transmission (transfer) time can be omitted aside data processing time (computing). His model was perfectly valid for the that-time implementation using vacuum tubes, and he clearly limited its range of validity, excluding later computing technologies (producing “too fast” computing elements) and biological computing (with the “unsound” condition that the transfer (conduction) time is much longer than the processing (synaptic+neuronal) time).

However, he did not provide another “procedure” for his computing model, which could be used in the general case, when his omission for the timing relations was not valid. Despite this hiatus, the classic computing is used today for both technological and biological computing, because no other paradigm existed. We noticed that the timing relations (which we described quantitatively as ‘dispersion’ in his terms) gradually changed, and we hypothesized that today they are causing issues such as tremendous heat production of modern processors, desperately low computing performance of big artificial neural networks and distributed processing (such as supercomputers), among others. We also hypothesized that the general computing model shall describe biological computing systems, where the timing relations are the exact opposite of the relations von Neumann assumed in his paradigm; and because of this, the phenomena predicted by the theory were already noticed and those measurement data can underpin our theoretical conclusions.

The role of time in the processes is vital (although it became clear only gradually during the research, even for the PI). von Neumann clearly described that a real computing comprises several *chained* elementary computations, i.e., one has to deliver the result from an output section of one operating unit to the input section of another operating unit. This is also valid if physically only one computing unit exists and it performs successive operations as one single adapting operating unit.

The first step was to develop the mathematics which can describe the temporal behavior of computing systems where the information transmission speed is finite (the classic computing science, by assuming zero information transmission time, in analogy with the classical physics, assumes immediate interaction). The analogy offered Minkowski’s famous four dimensional *space-time* coordinates into our attention, which was originally developed exactly for our goal: to describe phenomena occurring in systems with finite interaction speed. This mathematics is famous from theory of special relativity, where it is combined with Lorentz-transform to describe motion in moving reference frames of space-time. One difference is that our case is more simple: we have only one frame of reference (i.e., we need only the four-dimensional coordinates). Another difference is that we need to use another *scale factor*: we introduced *time-space* coordinates where the space coordinates (in this sense: the length of the signal path: wiring length in technology, axon length in biology) are divided by the interaction speed (speed of the electromagnetic interaction or conduction velocity); in this way describing events with *time-only* coordinates instead of *distance-only* coordinates we know from theory of special relativity. (This change was necessary because both in electronics and in biology the time is the relevant measurable parameter.) These four-dimensional coordinates are equivalent with the coordinates using another scale factor: the phenomena can be described with both coordinates, in an identical way. We extended Einstein's famous thought experiment with introducing "processing time" into the model and showed that the computing events can be described as simple four-vector operations. From theoretical point of view, we introduced a "temporal logic", where the value of a logical function related to an event depends also on time and place of the event. In this way on the one side we could keep the solid mathematical background of computing science, on the other side we could consider real implementations with finite signal propagation speed (published in journal Informatics).

We applied the theoretical conclusions to several technological ideas and implementations, inherited from the past where the temporal relations were quite different. In connection with checking elementary processor operations, we have pointed out that the "cost function" of processor design is outdated. At the beginning of computing, the design goal was to minimize the transistors in the circuits, despite that the different signal paths in the design have different temporal length. For today, the different internal signal arrival times led to unwanted flops in the internals of digital circuits, and are the primary reason of the enormously increased power consumption of modern processors. We pointed out that using *the time difference between the signal paths to a computing element as a cost function* would result in a much reduced energy consumption (published in "Foundations of Computer Science").

The cooperating processors must communicate in solving their joint task. The technical way as they cooperate (most commonly, the parallelized sequential processing) comprises inherent non-payload activity, which increases with the number of processors and becomes a theoretical *performance gain* limiting factor. We predicted that large supercomputers have theoretical absolute *performance gain* limit, and have published its value in J. Supercomputing. Our prediction still meets the experienced practical performance limits, experienced by the developers of supercomputers in different countries. Our results were also presented in the frame of a seminar of the Human Brain Project why the present architectures and principles are not feasible to simulate the brain:

https://www.fz-juelich.de/SharedDocs/Downloads/INM/INM-1/EN/Abstract_INM-1_Seminar_20210913.pdf

We explained why and how the *workload* (the program they run) affects the performance of supercomputers and why is it pointless to build vast (energy-wasting) "racing-only" supercomputers for solving practical tasks. We theoretically explained the experience, that for "real-life" tasks, after exceeding a critical (task- and architecture-dependent) number of cooperating processors, adding more processors to the system, the execution time starts even to increase rather than to decrease further. We also checked large and heavily communicating supercomputer applications (such as simulating the brain) and pointed out that with the present implementation technologies no more than a few hundred processors can be used for such tasks. In general, the vast supercomputers cannot use more than a few dozens of processors at the same "real-life" task, and even the benchmark programs cannot utilize all of their processors. We also explained (in Brain Informatics) why implementing the brain simulation in hardware faces the same performance gain roof-line as in software, furthermore why the present architectural principles disable simulating (at least larger portions of) the human brain. Based on the virtue of workload, we also explained why the artificial neural networks have such a low computing efficiency (in Neural Computing and Applications); our theoretical results were underpinned by published measurements performed on electronically implemented different commercial artificial neural networks. We applied our research results to the recently preferred cloud-type computing systems by discussing some typical use cases (in review in Parallel Computing).

Using the time-aware approach we analyzed the effect of the technical implementation of the bus connecting computing components and demonstrated that this component alone blocks the operation of vast systems. (We underpinned our theoretical conclusions with the results of the simulator; the simulator itself is not the subject of an independent publication.) The idea of having the single high-speed bus means at the same time the need for contending for the right of using that single resource. The arbitration time increases with the number of computing elements irrationally: at the end of the time of the computation, the last of the 9.4 million cores can send its result after 40 minutes waiting over the 200 Gbit/s single high speed bus. Using sequential buses alone prevents large neural networks to operate at a reasonable computing efficiency.

Our research in the field of biology was initially confined to finding phenomena our mathematics can describe, but we experienced that introducing the true temporal behavior into biology is a new idea. We fund that although the word "spatiotemporal behavior" was known and used in neuroscience, the meaning did not include that *spatial and temporal coordinates of the neural signals are connected* by the signal transmission speed. Probably because of the lack of appropriate mathematics, different approximations (considering the time as one of the independent coordinates) were in use and provided contradictory results. Our generalized computing model successfully described biological computing, too (in Acta Biotheoretica).

The fact that our hypothesis meets the experimental results of biological research encouraged us to construct an abstract neuron model which can describe the experienced behavior of biological neurons. The model (in terms of electronics) is a combo: its input section comprises essentially an analog electric condenser, with gated operation. Its synapses are input gated by its ion channels (the chemically delivered ions increase the potential which opens the ion gates) and are output gated by the membrane potential (reaching the threshold value closes the ion gates). The operation of its output section is triggered by the event of reaching the threshold potential. After triggering, preparing the spike becomes independent from the input signals of the neuron: it is defined by the biophysical parameters of the neuron (its operation is in resemblance with that of the digital circuits). We pointed out theoretically, that the membrane acts as an internal memory with time-dependent content (in accordance with the experiments): in addition to the contributions delivered through its synapses, this starting non-equilibrium potential also affects the time of reaching the threshold potential. The neural computation is performed in a time window which is opened by the first arriving spike and closed by reaching the threshold potential of the membrane. The result of the computation depends on the synaptic inputs and the internal memory; a memory-less automaton does not describe, in general, the operation of a neuron. (in Entropy, in review in Biology)

In the light of our results, we hypothesized that the neural computation is essentially manipulating the time appropriately. Storing information means fixing the time of sending information; learning means adjusting the time of information sending. We hypothesized two possible mechanisms for short-time and long-time learning, and demonstrated that they manifest in the same effect of delivering signals in a shorter time. We found published neurophysiological and anatomical evidence for those learning mechanisms. (in Acta Biotheoretica)

From our model we successfully derived theoretically the length of time of a firing cycle, the reciprocal of which is known as the momentary firing rate. We pointed out that the firing rate is not an appropriate measure of the neuronal communication: it comprises a "foreign contribution" (due to the internal memory of the neuron and a delay contribution from its network neighbors). Its statistical behavior shows that in addition to its mean and variance also parameters skewness and kurtosis must be hypothesized to describe its distribution. The dependence of the two latter parameters on the experimental conditions clearly shows that *two* temporal contributions (which depend on the experimental conditions) are present, and their effect is erroneously attributed to one contribution. Our results suggest that the operation of neurons cannot be understood without its network and vice versa. We also hypothesized that the experienced "skewed" distributions (the presence of which is considered important for the operation of the brain) can be the result of a simple experimental data handling artifact, but can also show an evidence for the importance of the dynamic processes in the brain (in Entropy).

After discovering the temporal behavior of the neural operation, we hypothesized that we need to reinterpret the vital virtues of "computing", "processing" and "information" (with all attached virtues such as entropy). We pointed out that the classical information theory (worked out for electric communication) *must not be used for neural communication*, simply because its required conditions of applicability are not met. This also means that the results and conclusions, derived using classic information theory outside of its range of validity, are wrong. Our conclusion is in line with that of the mathematicians. By scrutinizing our model, we pointed out that – in line with earlier guesses – the neural information processing is neither analog, nor digital, nor a mixture of the two modes of the man-made electronic systems. Its mode in "neural", which under certain conditions shows considerable resemblance to either of the two mentioned modes. We called the attention to that the too close parallels between biology and electronics are not only misleading, but they can also misguide both fields.

Essentially from the general computing model and the millions of times different interaction speed we successfully derived that in a communication, messages have two components. One of them only delivers information on the distance of the sender object to the receiver one (an unintended communication, a consequence of the finite interaction speed) while the other one is a signal which the sender collects to the receiver. We have shown that these two message components, although seen in both technological and biological communication, are handled differently by them. The technological computing dismisses (and suppresses, see clock distribution tree) the temporal part and uses only the signal part of it. In contrast, the biological computing cannot use the signal part (the spikes have identical form), so it must use the temporal part. However, the absolute time has no stand-alone meaning (except the length of the path), so the real information source is the (change of the) temporal distance of consecutive spikes. The theory shows that the two fields use two different approximations to the transferred information.

By applying our theoretical model to biology we pointed out that the brain really computes, although we need to interpret most related notions – including the representation, processing and transfer of information - in a drastically different way. The introduced neuronal model is compatible with all published data. We successfully reinterpreted neuronal entropy and communication bandwidth limitations. We pointed out that bandwidth of biological communication can depend also on biological factors (In Entropy, Acta Biotheoretica, in review in Biology).

These research results enabled us to make "intellectual excursions" to some related fields. After discussing the features of learning in biology, we compared the biological learning to the technological one. The comparison resulted in conclusion that learning and machine learning, furthermore intelligence and machine intelligence, are essentially different notions, and practically use only that same name (in J. Artificial Intelligence and Machine Learning). The research results also enabled us to analyze the popular idea of using super-quick components for computing. We showed that it is worth to develop and use much quicker (and much more expensive) elements, technological solutions, new materials, etc., only if the corresponding wiring can be can be reduced proportionally, too. Memristors (and memristor arrays) are frequently proposed, among others, to replace or emulate biological neurons. In a recent research paper (in review in J. Low Power Electronics and Applications) we analyzed the abstract models of neuron and memristor, and showed that they have very different features, so their use in biology-related applications is very limited. The analysis of the temporal operation also enabled us to propose (a conference paper in series of "Foundations of Computer Science") an architecture where the technological limitations (such as the single-bus interconnection, the single-core oriented thinking, different memory levels) approach much closely the biological ones.

Our results attracted the attention of book publishers, too. Presently we are working on a book, invited by Francis and Taylor, entitled "Beyond computing: the time in technological and biological computing". We can summarize that our research targeted a "white spot" of (both technological and biological) computing, and its success opened new research fields in (the widely interpreted) computing science, the time-aware computing.