

# Neumann hiányzó második tervezete (amit tartalmaznia kellene) von Neumann's missing "Second Draft": what it should contain

**VÉGH János**  
egyetemi tanár

Miskolci Egyetem, Vegh.Janos@gmail.com

## Abstract

*Computing science is based on a computing paradigm that is not valid anymore for today's technology. Von Neumann's model is perfect, but now it is used outside of its range of validity. The paper introduces a quantitative merit for dispersion, which is vital for computing performance and energy consumption, and demonstrates how its value increased. Dispersion limits the utility of new materials/effects/architectures if the related transfer time cannot be proportionally mitigated.*

**Keywords:** modern computing paradigm, performance limitation, efficiency, parallelized computing, supercomputing, high-performance computing, von Neumann architecture

## Kivonat

*A számítástechnika olyan számítási paradigmán alapul, amely a mai technológia esetében már nem érvényes. Bár a Neumann-modell tökéletes, alkalmazása jelenleg már túllép annak érvényességi tartományán. A tanulmány mennyiségi mérőszámot vezet be – amely kulcsfontosságú a számítási teljesítmény és az energiafogyasztás szempontjából –, és bemutatja, hogyan növekedett annak értéke. A diszperzió korlátozza az új anyagok, architektúrák és fizikai jelenségek hasznosíthatóságát is.*

**Kulcsszavak:** modern számítási elmélet, teljesítőképesség korlát, párhuzamosított számítás, nagy teljesítőképességű számítások, Neumann-architektúra

## 1. BEVEZETÉS

A mai számítástechnika alapján általában von Neumann híres, „First Draft” (Első tervezet) című munkáját [1] tekintik<sup>1</sup>. Ebben a jelentésben az agy neuronális működésével párhuzamba állítva tárgyalta saját paradigmájának elveit és technikai megvalósítását, így modellje agyi ihletésű volt. Kiemelte a számítási folyamat időzítési viszonyait, és elemezte a késleltetések szerepét a számítási láncban. Felhívta a figyelmet arra is, hogy az átviteli idő a számítási folyamat szerves része: legyen szó biológiai, digitális vagy analóg rendszerről, a feldolgozás nem kezdődhet meg, amíg a bemeneti operandusok nem érkeznek meg a művelet helyére, és fordítva: a kimeneti operandus sem továbbítható, amíg a számítási folyamat le nem zárul. A számítási láncban az átviteli idő és a műveleti idő kölcsönösen gátolják egymást. Mivel elektroncsöves megvalósítást tervezett, a modell egyszerűsítése érdekében javasolta az átviteli idő figyelmen kívül hagyását a feldolgozási időhöz képest. Más szóval, közelítése azon alapul, hogy „számítószervi” egységei között azonnali a kölcsönhatás. Kifejezetten figyelmeztetett azonban arra, hogy helytelen lenne ezt a klasszikus – az átviteli időt figyelmen kívül hagyó – paradigmát a neuronális működés leírására használni, mivel ott az ingerületvezetési idő hosszabb, mint a szinaptikus idő. Egy 300 m<sup>2</sup>-es számítógéptermet és a becsült 3000 vákuumcsövet alapul véve von Neumann a vákuumcsövek közötti 30 cm-es távolságot tekintette kritikus értéknek. Ezen a távolságon az átviteli idő mintegy három nagyságrenddel kisebb, mint a feldolgozási idő (az általa említett „lépések” között). E korlátok mellett indokolta von Neumann az eljárást kizárólag vákuumcsöves technológia esetében.

<sup>1</sup> Ezt a publikációt gyakran „von Neumann-architektúráként” emlegetik, bár von Neumann már az első mondatban tisztázza: „Az alábbi megfontolások egy rendkívül nagy sebességű, automatikus digitális számítógépes rendszer szerkezetével, és különösen annak logikai vezérlésével foglalkoznak.”

Egy másik tényező, amely – saját szavaival élve – *érvényteleníti* paradigmáját, az az eset, amikor a processzor „túl gyors”. Megjegyezte, hogy százszoros frekvencia (kb. MHz) alkalmazása esetén – még elektroncsövek használatakor is – érvénytelenné válik az általa javasolt elhanyagolás. Ilyen frekvencián az átviteli idő megközelíti (vagy akár meg is haladja) a feldolgozási idő nagyságrendjét, így az átviteli idő figyelmen kívül hagyása már nem indokolható: a *látszólagos feldolgozási idő* (az egymást követő számítási műveletek közötti órajel-idő) több tíz százalékkal tér el a *fizikai feldolgozási időtől*.

## 2. AZ IDŐBELI JELLEMZŐK TÖRTÉNETE

Abban az időben, vákuumcsövek használata mellett, nem volt reális ilyen frekvenciák elérése. Napjainkban a processzorchip mérete a 3 cm-es tartományba esik, és ennek megfelelően az átviteli idő elérheti az 1 ns-os nagyságrendet. A feldolgozási idő (abban az értelemben, ahogyan azt Neumann János használta) jóval az 1 ns-os tartomány alatt marad, így érdemes megvizsgálni, hogy a mai számítástechnika alapulhat-e még a klasszikus paradigmán. „Ezen új hardver [neuromorf számítástechnika] megépítése szükségessé teszi az elektronika újragondolását”; „több fizikára és anyagtudományra van szükség” [2]. Ez azonban nem elegendő. Legalább némi fizikára magában a számítási paradigmában is szükség van ahhoz, hogy kiaknázhassuk a kutatott új jelenségek és anyagok előnyeit. Megfigyelték, hogy „a számítógépek teljesítménye óriási fejlődésen ment keresztül az elmúlt 60 évben, ám ez a fejlődés jelentősen lelassult az utóbbi évtizedben az alapvető számítási primitívek korlátai miatt” [3]. Az igazi ok azonban egy szinttel mélyebben rejlik: az egyébként helyes számítási modellt olyan technológiai körülmények között alkalmazzuk, amelyek mellett a klasszikus paradigmában tett egyszerűsítések már nem érvényesek.

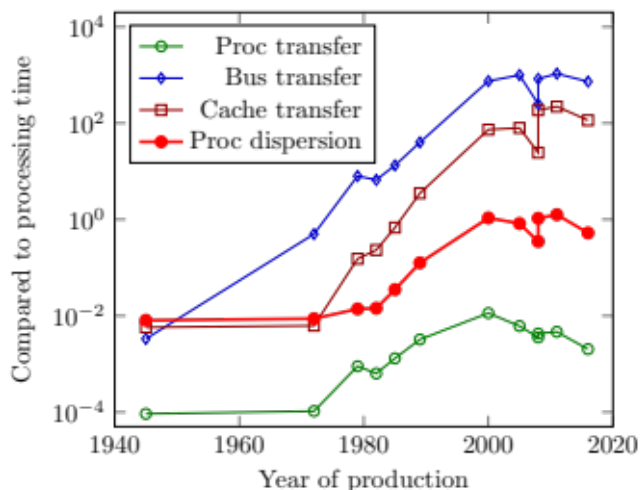
Mivel Neumann János az egyszerűsítéseit a *továbbítási időtől eltekintve* vezette le, az alábbiakban meghatározunk néhány mennyiségi jellemzőt. Az 1. ábra ezeknek a processzor gyártási évétől való függését mutatja. A műszaki adatok nyilvánosan hozzáférhető forrásokból<sup>2</sup> származnak. A jellemző értékek durva és némileg önkényes közelítések eredményei. Következetes alkalmazásuk azonban lehetővé teszi, hogy korlátozott érvényességű következtetéseket vonjunk le anélkül, hogy szükségünk lenne bizalmas technológiai adatokra. A feldolgozóelemek közötti távolságot kétféleképpen becsüljük meg. A vákuumcsöveknél fentebb ismertetett módszert alkalmazva kiszámítjuk a processzor területének és a tranzisztorok számának hányadosából vont négyzetgyököt. Ez az érték a tranzisztorok egyfajta „átlagos távolságát” adja meg. Ezt tekintjük annak a minimális távolságnak, amelyet a jeleknek a tranzisztorok között meg kell tenniük<sup>3</sup>. Ezt az értéket az 1. ábrán „Proc transfer”-ként (processzoron belüli átvitelként) tüntetjük fel. A maximális érték a chipen található két legtávolabbi feldolgozóelem közötti távolság<sup>4</sup>. Mivel a feldolgozóelemeket és a tárolóelemeket rendszerint külön technológiai blokkokként állítják elő, és vezetékekkel (ún. busszal) kötik össze őket, megbecsültük a „buszátviteli” időt is. Így a memória-hozzáférés ideje a buszátviteli idővel növekszik. E hatás miatt feltételeztük, hogy a gyorsítótár (cache) előnyösen helyezhető el a processzorméret felének megfelelő (átlagos) távolságban. Ezt az időt „Cache transfer”-ként (gyorsítótár-átviteli időként) jelöljük.

A gyorsítótárak az 1980-as évek végén jelentek meg. Nyilvánvalóvá vált, hogy a buszátvitel jelentősen megnöveli a memóriaátviteli időt. A gyorsítótár használata nagyságrenddel javíthatja a rendszerek teljesítményét (bár a gyorsítótárra vonatkozó adatok minden processzor esetében kiszámíthatók). Tekintettel arra, hogy „a hangsúly a diszperzió kiküszöbölésén van” [1], a technológiai adatok felhasználásával a szóródást a minimális és maximális „Proc transfer” idők mértani közepének és a feldolgozási időnek a hányadosaként határozzuk meg. Neumann János absztrakciójában a „jól meghatározott, szóródásmentes  $\tau$  szinaptikus késleltetés [feldolgozási idő]” szerepel.

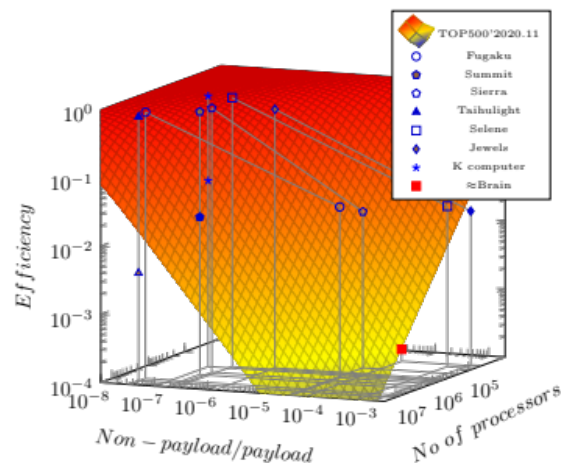
<sup>2</sup> [https://en.wikipedia.org/wiki/Transistor\\_count](https://en.wikipedia.org/wiki/Transistor_count)

<sup>3</sup> Megjegyzendő, hogy ez az átviteli idő szintén drasztikusan nő a tranzisztorok számával, önmagában azonban nem érvényteleníti a klasszikus paradigmát

<sup>4</sup> Nyilvánvaló, hogy az órajel-tartományok és a többmagos processzorok bevezetése árnyalja a képet. Bizalmas technológiai adatok hiányában azonban nem tudunk pontosabb becslést adni



1. ábra: A számítógép processzorok időjellemzői a gyártási év függvényében. Sok tranzisztor becsúfolása rontotta a jellemzőket.



2. ábra: A felület a szuperszámítógépek teljesítő-képességének elvi határa „versenylő” (HPL) és „igásló” (HPCG) üzemmódban, a legjobbak mért értékeivel.

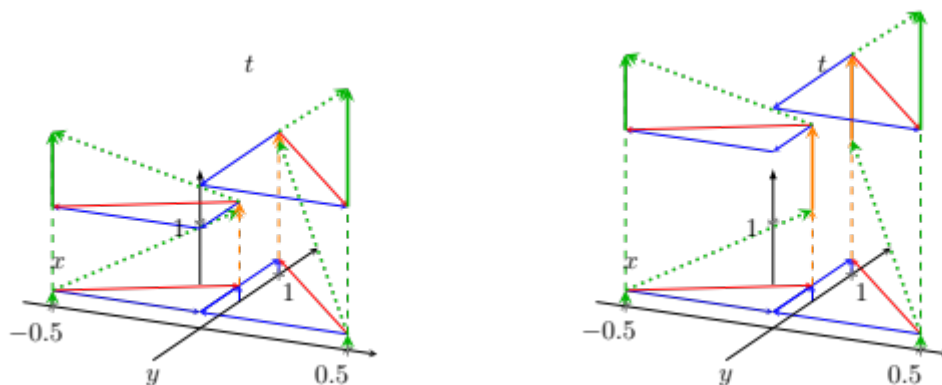
A mi definíciónk szerint az EDVAC szórása (legfeljebb) 1%-os; ezzel indokolta eljárását Neumann János. Érdekes párhuzam, hogy az EDVAC és az Intel 8008 ugyanannyi feldolgozóelemmel rendelkezik. A processzor és a gyorsítótár közötti relatív átviteli idők nagyságrendje megegyezik. Megjegyzendő azonban, hogy a buszátviteli idő jelentősége megnőtt, és a személyi számítógépek esetében meghatározóvá vált az egy-magos (single-processor) teljesítmény szempontjából. Egy évtizeddel később a busz fizikai mérete tette szükségessé a gyorsítótárak bevezetését. A fizikai méret az összes relatív átviteli idő telítődéséhez vezetett<sup>5</sup>. Az elmúlt évek relatív időértékeiben tapasztalt enyhe csökkenés valószínűleg annak tudható be, hogy számítási módszerünk érzékeny a többmagos (multi-core) architektúrák elterjedésére; ez arra utal, hogy érdemes lenne megismételni az elemzést saját (gyártói) technológiai adatok felhasználásával. Ahogy a „Proc dispersion” (processzor-szórás) diagram görbéje mutatja, a mai technológiában a szórás mértéke közel áll az egységhez. Ez a nagyfokú szórás azt jelenti, hogy a mai processzorok működési módja közelebb áll agyunk működési módjához, mint a klasszikus paradigmában absztrahált működési modellhez. Agyunk esetében azonban kifejezett „tér-idő” jellegű viselkedést feltételezünk, így leírására a klasszikus paradigma alkalmazása „helytelen”. Vagyis már nem alkalmazhatjuk a „szórásmentes” klasszikus paradigmát<sup>6</sup>. Ez a magas szórásérték az oka annak, hogy a bemeneti teljesítménynek csak kis töredéke hasznosítható számításra; a többi disszipálódik (hővé alakul). Ezt azóta tapasztaljuk, amióta a szórás mértéke mintegy két évtizeddel ezelőtt megközelítette az egységet. Ugyanabban a szakaszban von Neumann így fogalmazott: „Javasoljuk, hogy a  $\tau$  késleltetéseket abszolút időegységekként használjuk, amelyekre támaszkodva szinkronizálhatjuk az eszköz különböző részeinek működését. Egy ilyen elrendezés előnyei azonnal nyilvánvalóak.” Állítása valóban érvényes volt az általa feltételezett, jól meghatározott, diszperziómentes  $\tau$  szinaptikus késleltetésre, ám a mai processzorok esetében ez már egyáltalán nem állja meg a helyét. Az aszinkron működési módok vizsgálatára irányuló közelmúltbeli törekvéseket [2,3] az a felismerés motiválja, hogy a jelenlegi szinkron működési mód hátrányos a nem diszperziómentes világban. Saját korában, a vákuumcsöves technológia idején von Neumann nem érezte szükségét annak, hogy megvitassa, milyen eljárás indokolhatja a számítási művelet leírását nem diszperziómentes esetben. Ugyanakkor javasolta, hogy bármely új, jövőbeli technológia esetében vizsgáljuk felül a paradigmájában alkalmazott egyszerűsítések (elhanyagolások) érvényességét. Ahhoz, hogy a jelenlegi technológiák számára szilárd számítási paradigmát alakíthassunk ki, figyelembe kell vennünk az átviteli idő és a feldolgozási idő arányát; ezt már nem hagyhatjuk figyelmen kívül. A valódi kérdés – amelynek tárgyalása hiányzik az „Első tervezetből” – az, hogy milyen eljárást kell követni, ha az átviteli idő nem elhanyagolható?

<sup>5</sup> A „Moore-korszak vége” valódi oka az, hogy Moore megfigyelése nem érvényes a fizikai buszméretre.

<sup>6</sup> A diagramgörbék platójának elérése egybeesik az „explicit párhuzamos utasításkészletű számítógép” (EPIC) bevezetésével [4]: ez volt a klasszikus paradigma által lehetővé tett maximum.

### 3. AZ IDŐBELI LOGIKA BEVEZETÉSE A SZÁMÍTÁSTECHNIKÁBA

Bár Neumann János kifejezetten említette, hogy az elektromágneses hullámok terjedési sebessége korlátozza az elektronikus alkatrészek működési sebességét, ezt a hatást a számítástechnikában egészen a közel-múltig nem vették figyelembe. Ezzel szemben a biológiában már nagyon korán felismerték a „tér-idő” jellegű viselkedést. Napjainkban jellemző törekvés az idegrendszeri működés elméleti leírása és elektronikus modellezése a Neumann által javasolt számítástechnikai paradigma alapján – amely azonban kétségtelenül nem érvényes a mai technológiára. Neumann szerint kétszeresen is *hibás* megközelítés az idegrendszeri működés utánzására olyan paradigmát használni, amely eleve *alkalmatlan* erre a célra, ráadásul olyan technológiai alapokon (a vákuumcsöveken kívül), amelyek magát a paradigmát is *érvénytelenítik*. Szerencsére a tér-idő jellegű viselkedés olyan „eljárást” kínál, amely akkor is alkalmazható, ha az átviteli idő hosszabb, mint a feldolgozási idő. Bár a „tér-idő” elnevezést használják, a biológia rendszereinek viselkedését elkülönült tér- és időfüggvényekkel írják le (ami miatt eseti jellegű megoldásokra és feltevésekre van szükség az egyes problémák kezelésénél). Egy közös vonásuk azonban mégis van a műszaki számítástechnikával: mindkét esetben korlátozott az információátvitel sebessége. Első ránézésre furcsának tűnhet az ilyen rendszerek leírása a Minkowski-formalizmussal (illetve annak inverzével), mivel az az Einstein-féle speciális relativitáselmélet kapcsán vált ismertté. Eredeti formájában azonban csupán egy határsebesség létezését tételezi fel. Ahogy azt a [5] tanulmány tárgyalja, ez a tulajdonság lehetővé teszi számunkra, hogy leírjuk az információfeldolgozás helyes viselkedését mind a tudományos alapú műszaki megvalósításokban, mind a biológiában, az átviteli és feldolgozási idő tetszőleges kombinációja esetén. A kulcs gondolat az, hogy a számítástechnikai komponensek (legyenek azok Si-kapuk, processzormagok, hálózati csomópontok, valamint biológiai vagy mesterséges neuronok) közötti térbeli távolságokat időmértékre váltsuk át (a jelút mentén érvényes határsebesség alapján mérve). Ilyen ábrázolást mutat a 3. ábra. A látszólag három-dimenziós ábrán a processzorok az  $x$ , a memóriák az  $y$  tengely mentén vannak elhelyezve. A két processzormag adatot kér a memóriától, majd azzal műveletet végez. A narancsszínű függőleges nyilak ábrázolják a fizikai gyorsítótár működési idejét, a zöld függőleges nyilak pedig a látszólagos elérési időt. A szaggatott nyilak azt hangsúlyozzák, hogy a művelet végzéséhez az adatokat (címet és memória tartalmát) szállítani kell a művelet végrehajtó processzor és a memória között.



3. ábra: A processzor chipen belül elhelyezett cache memória teljesítménye különböző működési idők esetén, azonos geometriában. Az  $x=-0.5$  és  $x=0.5$  időtávolságokban elhelyezett processzor magok  $y=0.5$  és  $y=1.0$  időpontokban érik el a gyorsítótárat. A jobb oldali ábrán a gyorsítótár működési ideje tízszer kisebb, a látszólagos elérési idő azonban csak kétszer kisebb.

A látszólagos elérési idő sokkal érzékenyebb a gyorsítótár helyzetére mint annak sebességére.

### 4. AZ IDŐBELI VISELKEDÉSBŐL ADÓDÓ JELENSÉGEK

A számítástechnikában „több fizikára... van szükség” [2]. Az időbeli működés leírása kiváló példa arra, hogy a soronkénti fordítás (szekvenciális programozás, más néven Neumann-típusú programozás [6]) formálisan csak logikai függőséget vezet be, ám technikai megvalósítása révén implicit és szükségszerű módon időbeli viselkedést is eredményez. Az időbeli viselkedés figyelembevétele hosszabb késleltetést eredményez a mai technológiában használt kapukból felépített összeadó esetében, mint a Neumann által feltételezett logikai összeadóé. Az összetevők időbeli viselkedése drasztikus teljesítménycsökkenést eredményez párhuzamosított számítások esetén. Érdeemes felidézni, hogy „ez a teljesítménycsökkenés nem az architektúra hibája, hanem a

korlátozott párhuzamosíthatóság következménye” [7]. Már Neumann felhívta a figyelmet: míg egy kapu nem kapja meg az összes bemeneti jelét, a kimenete határozatlan. E jelenség drasztikus következményeivel kapcsolatban lásd a szuperszámítógépek teljesítményére vonatkozó 2. ábrát, továbbá a mesterséges neurális hálózatok tanítására [8] és nagy feladatok (pl. agy szimulálás [9]) gyakorolt hatást. A teljesítménycsökkenés olyan erőteljes, hogy a nagy kihívást jelentő feladatok megoldásához 10-11 nagyságrendet(!) kellene javulni a hatékonyságnak [10].

Időbeli viselkedésük korlátozza az új anyagok és jelenségek alkalmazhatóságát is; ez egy rendkívül népszerű kutatási irány, különösen a neuronális műveletek terén [2]. Bár a látszólagos működés felgyorsul, amikor tízszer gyorsabb (és általában százszor drágább) gyorsítótárat használunk, az elért sebességnövekedés nem arányos a gyorsítótár fizikai működési sebességének növekedésével, lásd 3. ábra. Hasonlóképpen gazdaságtalan a processzor műveleti sebességének növelése, ha nem tudjuk az adatátvitel sebességét arányosan növelni.

## 5. ÖSSZEFOGLALÁS

A jelenleg széles körben alkalmazott klasszikus számítási paradigma – feltalálója szerint – kizárólag elektroncsövekre (és azok időzírti viszonyaira) érvényes. A rendszerkomponensek közötti azonnali kölcsönhatást feltételezi, ami elméletileg hibás, és ellentmond a mindennapi tapasztalatoknak. A jelenségek skálája széles: az elakadt egyprocesszoros teljesítménytől (amelyet inkább fűtésre, mintsem számításra használnak) kezdve – a saját belső teljesítménykorlátaikat átlépni képtelen szuperszámítógépeken át, amelyek valós feladatok esetén a *névleges teljesítményük* mindössze kb. 1%-át kitevő *hasznos teljesítményt* nyújtanak – egészen a mesterséges neurális hálózatok skálázódásának elakadásáig. „Az új hardver megépítése szükségessé teszi az elektronika újragondolását” [2]. Ebben a helyzetben két lehetőségünk van. Kiegészíthetjük a tudományág és folyóiratainak jelenlegi elnevezését úgy, hogy az tükrözze: „Számítástechnika kizárólag elektroncsövekhez”. Továbbá indíthatunk egy új tudományágot „Modern számítástechnika” néven (párhuzamban a klasszikus és modern tudomány közötti különbséggel), amely az időbeli logikára épül, és a legkorszerűbb technológián alapuló számítástechnikát írja le. Így léphetünk a „következő szintre”. Még a számítástechnika „újraindítása” sem lehetséges ilyen drasztikus változtatás nélkül. A másik lehetőség a klasszikus számítástudomány megtartása és annak elismerése, hogy „Game over” „lejárt az időnk” [11].

## IRODALMI HIVATKOZÁSOK

- [1] Neumann, J.: First draft of a report on the EDVAC. IEEE Annals of the History of Computing, 15/4(1993)27-75, doi=10.1109/85.238389, <https://archive.org/details/vnedvac/>
- [2] Markovic, D., Mizrahi, A., Querlioz, D., Grollier, J.: Physics for neuromorphic computing, Nature Reviews Physics, 2(2020),499–510, <https://www.nature.com/articles/s42254-020-0208-2.pdf>
- [3] Mehonic, A., Kenyon, A. J.: Brain-inspired computing needs a master plan. Nature 604/7905 (2022)255-260. 10.1038/s41586-021-04362-w
- [4] Schlansker, M.S., Rau, B.R.: EPIC: Explicitly Parallel Instruction Computing. Computer 33/2(2000)37-45.
- [5] Végh, J., Berki, Á. J.: On the spatiotemporal behavior in biology-mimicking computing systems. Acta Biotheoretica 70/4(2022)26 <https://link.springer.com/article/10.1007/s10441-022-09450-6>
- [6] Backus, J.: Can Programming Languages Be liberated from the von Neumann Style? A Functional Style and its Algebra of Programs. Communications of the ACM 21(1978)613-641
- [7] Singh, J.P., Hennessy, J.L., Gupta, A.: Scaling Parallel Programs for Multiprocessors: Methodology and Examples. Computer 26/7(1993)42-50 10.1109/MC.1993.274941
- [8] Végh, J.: Which scaling rule applies to Artificial Neural Networks. Neural Computing and Applications 33/1(2021) 16847–16864. 10.1007/s00521-021-06456-y <http://link.springer.com/article/10.1007/s00521-021-06456-y>
- [9] Végh, J.: How Amdahl's Law limits performance of large artificial neural networks. Brain Informatics 6/4(2019)1-11 10.1186/s40708-019-0097-2
- [10] Conklin, Alexander A., Kumar, Suhas: Solving the big computing problems in the twenty-first century. (2023) <https://www.nature.com/articles/s41928-023-00985-1.epdf>.
- [11] US National Research Council: The Future of Computing Performance: Game Over or Next Level? 2011 <http://science.energy.gov/~media/ascr/ascac/pdf/meetings/mar11/Yelick.pdf>