The role of time in computing

János Végh

Jülich, September 27, 2021

V0.05@2020.09.23 (https://github.com/jvegh/MultEdu V0.5.9)

Introduction: Cooperation in many-discipline projects

Blind monks examining an elephant

Q1A: Can we achieve EFlops performance?

Fugaku Holds Top Spot, Exascale Remains Elusive

June 28, 2021

TOP500 NEWS

GREEN500: Trend of steady progress with no big step toward newer technologies.

June 28, 2021

Although there was a trend of steady progress in the Green500, nothing has indicated a big step toward newer technologies.

The system to snag the No. 1 spot for the Green500 was MN-3 from Preferred Networks in Japan. Knocked from the top of the last list by NVIDIA DGX SuperPOD in the US, MN-3 is

SPONSORED ARTICLE

PRACE Software Strategy for European Exascale Systems

Sept. 1, 2021

Building on the successful implementation of the Partnership for Advanced Computing in Europe (PRACE), the European Commission (EC) has increased its efforts to develop a world-class supercomputing ecosystem in Europe. The EC, EuroHPC Joint Undertaking (JU) and EU Member States have made significant investments in European petascale and pre-exascale infrastructure, have put exascale supercomputers on the roadmap, and are actively exploring new post-exascale architectures. The return on investment will be directly linked

TOP500 News

Who are closer to Exascale, say it is unreachable

Who are farther, prepare for post-exascale

Q1B: What ingredients are needed to achieve it?

"the largest HPC hardware systems today contain more than 1 million cores, and exascale supercomputers with tens or hundreds of millions of cores will begin to arrive during the period 2020-2022" and "With a differentiated strategy and sufficient investment and political will, Europe can be a global player in HPC".

Europe Union's Action plan, 2016)

Q1B: What ingredients are needed to achieve it?

> *"the largest HPC hardware systems today contain more than 1 million cores, and exascale supercomputers with tens or hundreds of millions of cores will begin to arrive during the period 2020-2022" and "With a differentiated strategy and sufficient investment and political will, Europe can be a global player in HPC".*
>
> Europe Union's Action plan, 2016)

it "*can be seen in our current situation where the historical ten-year cadence between the attainment of megaflops, teraflops, and petaflops has not been the case for exaflops*"[1]

[1] Feldman M (2019) Exascale is not your grandfather's HPC. https://www.nextplatform.com/2019/10/22/exascale-is-not-your-grandfathers-hpc/

Q2A: What performance are we speaking about?

Chaotic news, communications, flyers, data sheets, benchmarks

- An (unexplained) 'empirical efficiency' exist
- Mismatching is payload/nominal performance
- Different operand lengths and representations
- Distributed systems have different efficiencies[2]
- (The efficiency depends on workload?)
- No role is devoted to systems' workloads
- Is there a theoretical upper limit for performance?
- (Is out of scope of feasibility studies, everywhere)

[2] https://spectrum.ieee.org/tech-talk/computing/hardware/two-different-top500-supercomputing-benchmarks-show-two-different-top-supercomputers

Q2A: What performance are we speaking about?

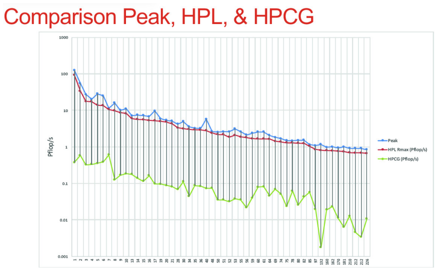
Chaotic news, communications, flyers, data sheets, benchmarks

- An (unexplained) 'empirical efficiency' exist
- Mismatching is payload/nominal performance
- Different operand lengths and representations
- Distributed systems have different efficiencies[2]
- (The efficiency depends on workload?)
- No role is devoted to systems' workloads
- Is there a theoretical upper limit for performance?
- (Is out of scope of feasibility studies, everywhere) Except here

[2] https://spectrum.ieee.org/tech-talk/computing/hardware/two-different-top500-supercomputing-benchmarks-show-two-different-top-supercomputers

Q2B: What happened to workload-dependent performance?

Different benchmark ⇒ different efficiency



Performance ratio, a few years ago

#	T	Site	Manufacturer	Computer	Country	HPCG [Pflop/s]	HPL [Pflop/s]	HPCG/ Peak	HPCG/ HPL
1	1	RIKEN-CCS	Fujitsu	Fugaku Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D	Japan	16.005	442.0	3.0%	3.6%
2	2	Oak Ridge National Laboratory	IBM	Summit IBM Power System, P9 22C 3.07 GHz, Volta GV100, EDR	USA	2.926	148.6	1.5%	2.0%
3	5	NERSC - Lawrence Berkeley National Laboratory	HPE	Perlmutter HPE Cray EX235n, AMD EPYC 64C 2.45GHz, NVIDIA A100, Slingshot-10	USA	1.905	64.6	2.0%	3.0%
4	3	Lawrence Livermore National Laboratory	IBM	Sierra IBM Power System, P9 22C 3.1 GHz, Volta GV100, EDR	USA	1.796	94.6	1.4%	1.9%
5	6	NVIDIA Corporation	NVIDIA	Selene DGX A100 SuperPOD, AMD 64C 2.25GHz, NVIDIA A100, Mellanox HDR	USA	1.623	63.5	2.1%	2.6%
6	8	Forschungszentrum Jülich (FZJ)	Atos	JUWELS Booster Module BullSequana XH2000, AMD EPYC 24C 2.8GHz, NVIDIA A100, Mell. HDR	Germany	1.275	44.1	1.8%	2.9%
7	11	Saudi Aramco	HPE	Dammam-7 Cray CS-Storm, Xeon 20C 2.5GHz, NVIDIA T. V100, IB HDR 100	Saudi Arabia	0.881	22.4	1.6%	3.9%
8	9	Eni S.p.A	Dell EMC	HPC5 PowerEdge C4140, Xeon 24C 2.1GHz, NVIDIA V100, Mellanox HDR	Italy	0.860	35.5	1.7%	2.4%
9	13	Information Technology Center, The University of Tokyo	Fujitsu	Wisteria/BDEC-01 (Odyssey) PRIMEHPC FX1000, A64FX 48C 2.2GHz, Tofu interconnect D	Japan	0.818	2[illegible]	[illegible]2%	3.7%
10	40	Japan Agency for Marine-Earth Science and Technology	NEC	Earth Simulator - SX-Aurora TSUBASA SX-Aurora TSUBASA A401-8, Vector Eng. Type20B 8C 1.6GHz, IB HDR 200	Japan	0.748	10[illegible]	[illegible]6%	7.5%

Performance ratio, today

A few years ago, the HPCG to HPL performance ratio was 0.2 %. . . 0.4%; today it is 3%. . . 7%

At the same computers. What happened?

New measurement method, algorithm, architecture; miracle; cheating?

Q3: Can we simulate the brain?

The ever larger computing systems inspire attempting simulating the operation of (larger parts of) the brain
Nature 571, S9 (2019): The four biggest challenges in brain simulation

- simulating the human brain would push the limits of even the exascale computers
- a biologically faithful simulation of the brain would require an almost limitless set of parameters
- no present technology can run large-scale simulations faster than in real time
- we lack a strong theory of how the brain works

Q4: Can we simulate a neuron?

Peter Sterling and Simon Laughlin: Principles of Neural Design

- "the broad definition of learning: use present information to adjust a circuit, to improve future performance"
- "we should not seek a special organ for 'information storage' — it is stored as it should be, in every circuit"
- "stored and retrieved directly"
- "Send only information that is needed, and send it as slowly as possible"

James V. Stone:Principles of Neural Information Theory

- "the more precisely spike timing is measured, the more information is gained"
- the information is mainly encoded in the 'temporal precision'
- time handling (Not time stamping!) is missing from simulators
- using information theory for neural information transfer needs revision

The Exascale Challenge

- There is a lot of vagueness in the exascale discussion and what it means to reach exascale. Let me propose a concrete and measurable goal.
- Build a system before 2020 that will be #1 on the TOP500 list with an Rmax > 1 exaflops
- Personal bet (with , Jülich, Germany) that we will not reach this goal by November 2019 (for $2,000 or €2000)
- Unfortunately, I think that I will win the bet. But I would rather see an exaflops before the end of the decade and loose my bet.

17

https://www.researchgate.net/publication/261879110_Why_we_need_Exascale_and_why_we_won't_get_there_by_2020

- Why you cannot win your bet, is a "conceptual flaw", rather than "technological imperfectness"
- We need to revise the very basics of computing
- We need to re-discuss the model of computing
- We need to upgrade our classic computing paradigm, for the current technology

AI, HPC, Processor: something new (different) needed

- DOE AI Expert Says New HPC Architecture Is Needed[3]
- "It's all about spatial delays and the tyranny of distance."
- "And the thing about the power there is we just get rid of all the transistors."
- There's always this and "then a magic compiler makes it all work right." And you, you just described the most magical of compilers I have ever heard about.
  Answer: [Laughter]. [4]

They are right, but there is no idea around

[3] https://www.nextplatform.com/2021/08/26/doe-ai-expert-says-new-hpc-architecture-is-needed/
[4] https://www.nextplatform.com/2021/07/12/gutting-decades-of-architecture-to-build-a-new-kind-of-processor/

"the vendors cannot test a system of this scale before it's deployed so we will have to trailblaze at deployment time to work out the kinks and make sure the system is stable.[5]"
The future is not brighter: "If you were expecting to get a big reveal of the TPUv4 architecture from the search engine giant and machine learning pioneer at its Google I/O 2021 conference this week, you were no doubt, like us, sorely disappointed.[6]"

[5] https://www.nextplatform.com/2021/05/21/first-look-at-frontier-supercomputers-storage-infrastructure/
[6] https://www.nextplatform.com/2021/05/21/google-hints-about-its-homegrown-tpuv4-ai-engines/

Scale is always a problem ...

"the vendors cannot test a system of this scale before it's deployed so we will have to trailblaze at deployment time to work out the kinks and make sure the system is stable.[5]"

The future is not brighter: "If you were expecting to get a big reveal of the TPUv4 architecture from the search engine giant and machine learning pioneer at its Google I/O 2021 conference this week, you were no doubt, like us, sorely disappointed.[6]"

The real issue is the incomplete understanding how computing works.

The scaling in strongly non-linear, because of physical reasons

[5] https://www.nextplatform.com/2021/05/21/first-look-at-frontier-supercomputers-storage-infrastructure/

[6] https://www.nextplatform.com/2021/05/21/google-hints-about-its-homegrown-tpuv4-ai-engines/

Just as early buildings and bridges sometimes fell to the ground—in unforeseen ways and with tragic consequences—many of our early societal-scale inference-and-decision-making systems are already exposing serious conceptual flaw

Michael I. Jordan: Jul 02, 2019

Artificial Intelligence—The Revolution Hasn't Happened Yet https://hdsr.mitpress.mit.edu/pub/wot7mkc1/release/9

Just as early buildings and bridges sometimes fell to the ground—in unforeseen ways and with tragic consequences—many of our early societal-scale inference-and-decision-making systems are already exposing serious conceptual flaws.

Michael I. Jordan: Jul 02, 2019

Artificial Intelligence—The Revolution Hasn't Happened Yet https://hdsr.mitpress.mit.edu/pub/wot7mkc1/release/9

Recall Aurora'17, Gyoukou, Aurora'21, Fugaku 40%
PEZY, Intel

Time: Standpoint of von Neumann

"*we will base our considerations on a hypothetical element, which functions essentially like a vacuum tube. . . . We reemphasize: This situation is only temporary, only a transient standpoint . . . After the conclusions of the preliminary discussion the elements will have to be reconsidered*" [7,8]

[7] J. Neumann: "First Draft of a Report on the EDVAC"
[8] J. Végh: "von Neumann's missing 'Second Draft': what it should contain"

Standpoint of von Neumann

"*we will base our considerations on a hypothetical element, which functions essentially like a vacuum tube. ... We reemphasize: This situation is only temporary, only a transient standpoint ... After the conclusions of the preliminary discussion* ***the elements will have to be reconsidered***" [7,8]

In his age, the processing time was in the *msec*, and the transfer time in the μsec region.

For today, the **ratio of times has reversed**

Since the beginning, technology passed several eras

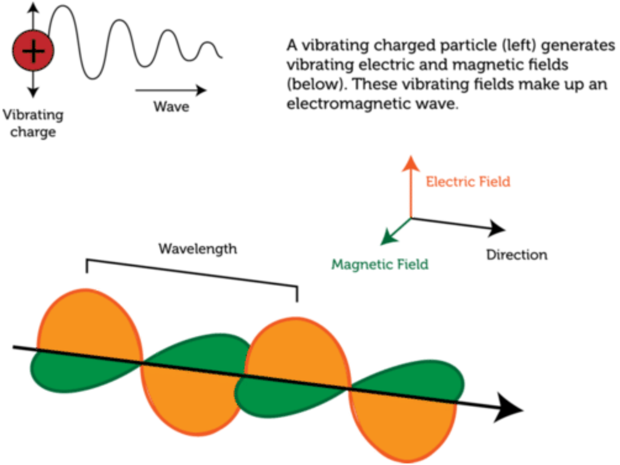
Computing paradigm did not change in the past three-quarter century

It is high time to "reconsider", and to include **time in computing science**

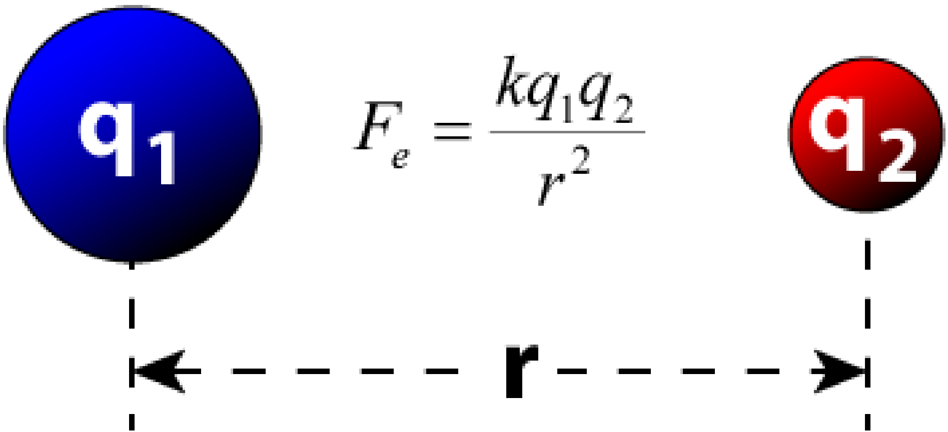
[7] J. Neumann: "First Draft of a Report on the EDVAC"
[8] J. Végh: "von Neumann's missing 'Second Draft': what it should contain"

"Time is not the subject of computing science"

Galilei at his trial



Creating electromagnetic waves is easy, explained at K-12 level

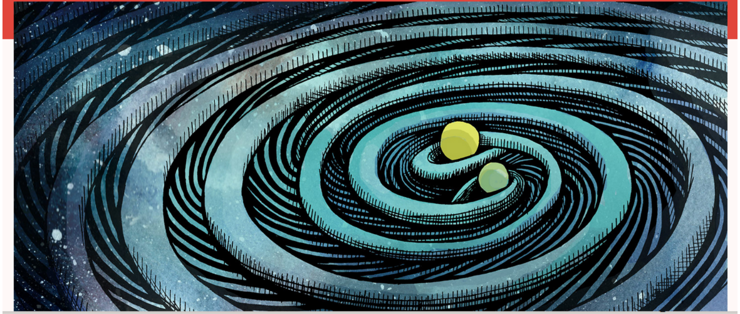


The Law does not have time-dependence, interaction speed cannot be derived

Nobel recognizes gravitational wave discovery

10/03/17 | By Kathryn Jepsen

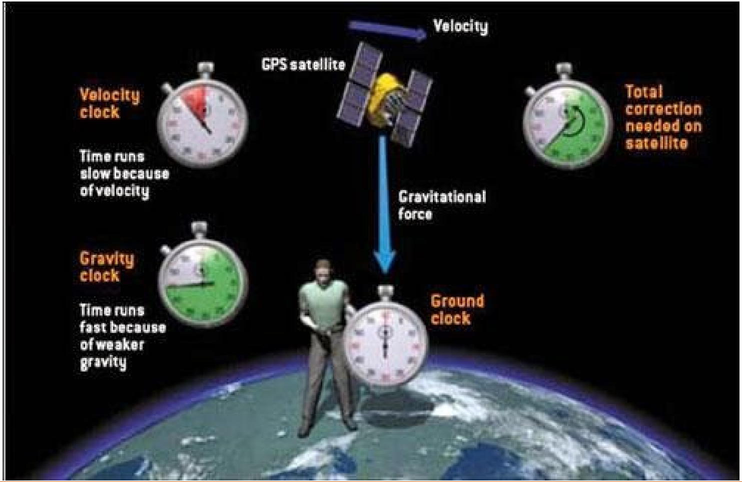
Scientists Rainer Weiss, Kip Thorne and Barry Barish won the 2017 Nobel Prize in Physics for their roles in creating the LIGO experiment.



The merit of (introducing) time

Is there any practical use of temporal behavior?

©http://unbonmotgroundswell.blogspot.com/2014/11/gps-and-relativity.html



Without its 38 μs relativistic correction, GPS can calculate position 100-fold worse computing systems "*have a clock signal which is distributed in a tree-like fashion into every tiny part of the chip. . . Approximately 30 % of the total consumption of a modern microprocessor is solely used for the clock signal distribution.*[9]

[9] Rainer Waser, Advanced Electronics Materials and Novel Devices (2012), isbn = 9783-527-40927-3

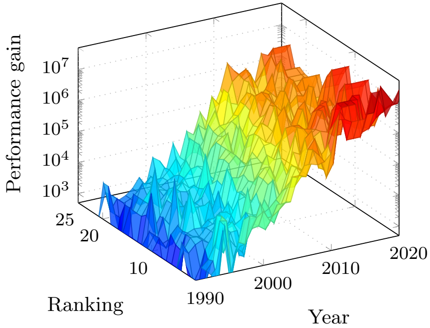
computing systems "*have a clock signal which is distributed in a tree-like fashion into every tiny part of the chip. . . Approximately 30 % of the total consumption of a modern microprocessor is solely used for the clock signal distribution.*[9]

i.e. 100% gas pedal plus 98% brake pedal

heating (and cooling!) rather than computing

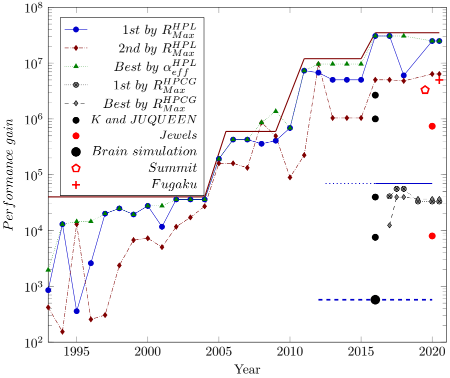
we must learn how to drive our new and expensive car

[9]Rainer Waser, Advanced Electronics Materials and Novel Devices (2012), isbn = 9783-527-40927-3



Performance gain development

- Gordon Bell prize for performance gain
- Gain of 1st 25 computers
- Using Linpack (HPL) enables us to compare data over decades
- Technology independent
- Two plateaus
- One hillside
- Some "bumps"



Performance gain development

- Data points derived from TOP500
- Workflows (HPL, HPCG, Brain)
- Different rooflines
- Technological ages
- Interconnection technology
- Computing principles
- Accelerators

Performance of parallelized sequential (distributed) systems

$$P(N,\alpha) = P_{single} \cdot N \quad (Gustafson)$$

Performance of parallelized sequential (distributed) systems

$$P(N,\alpha) = P_{single} \cdot N \cdot \boxed{\frac{1}{N\cdot(1-\alpha)+\alpha}} \quad (Amdahl)$$

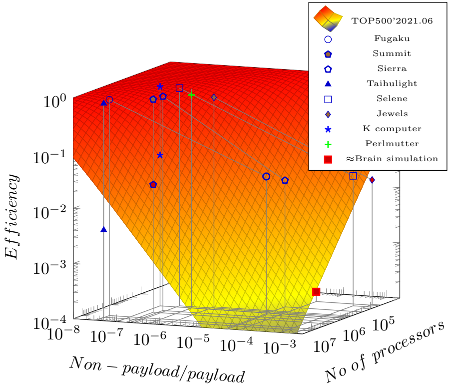
Performance of parallelized sequential (distributed) systems

$$P(N,\alpha) = P_{single} \cdot N \cdot \boxed{\frac{1}{N\cdot(1-\alpha)+\alpha}} \quad (Amdahl)$$

- α is the "goodness of parallelization"
- $(1-\alpha)$ is the non-payload performance, technological and scientific
- contributions to $(1-\alpha)_{theor}$ can be theoretically estimated
- $(1-\alpha)_{exp}$ can be derived from computing experiments
- $1/(1-\alpha)$ defines performance gain (speedup)

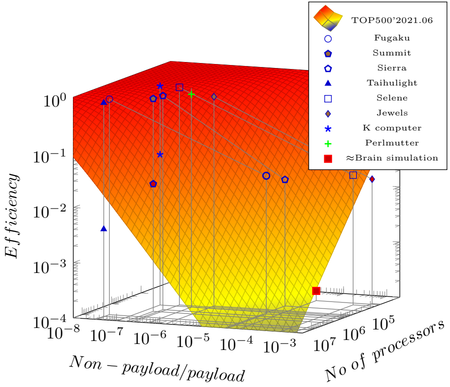
Analyzis method

- deriving $(1-\alpha)_{exp}$ from TOP500
- deriving $(1-\alpha)_{theor}^{contributions}$ from technological data
- validating method on data from TOP500
- change virtually N, predicting performance based on $(1-\alpha)_{exp}$
- first-order approximation (α constant)
- deriving performance limits based on $(1-\alpha)_{theor}^{contributions}$
- $1/(1-\alpha)$ defines performance gain (speedup)



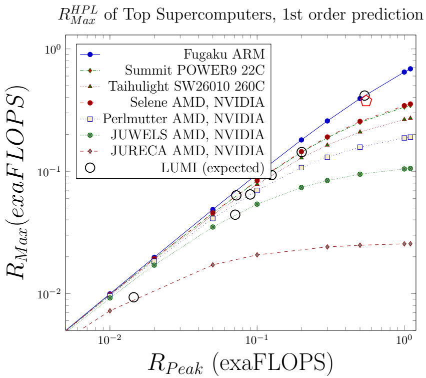
Payload efficiency of distributed system

- Efficiency depends also on the number of cores
- HPL and HPCG data are displayed
- Accelerators, etc. increase access time ⇒ decrease efficiency



Payload efficiency of distributed system

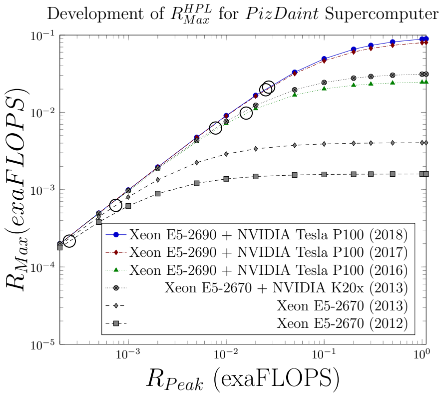
- Unaccelerated computers show the highest efficiency
- Higher workload sharply decreases computing efficiency
- Brain can be simulated at very low efficiency
- GPU accelerated supercomputer cannot use all their cores for solving real-life tasks

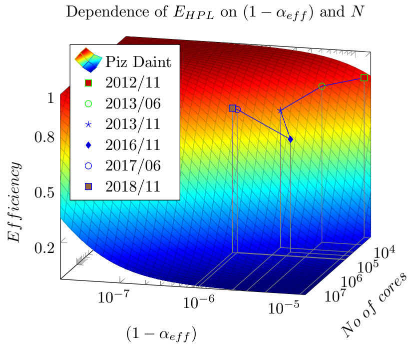


Payload efficiency of distributed system

- The first order approximation shows payload performance increase
- All configurations stalled below 0.5 EFlops (payload)
- The ones targeting above 1 EFlops (nominal) failed
- The highest payload performance is achieved without acceleration
- Using GPU limits payload performance

https://www.goodnewsfinland.com/world-leading-supercomputer-lumi-to-be-built-in-finland/

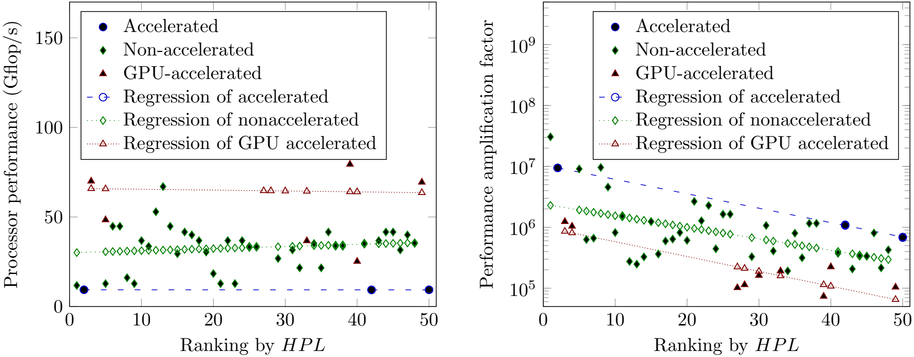




Payload performance of Piz Daint

The corresponding efficiency dependence

Different, well documented stages of development

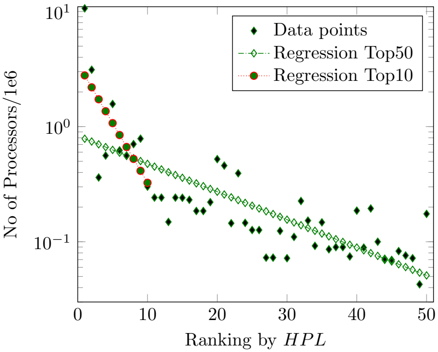


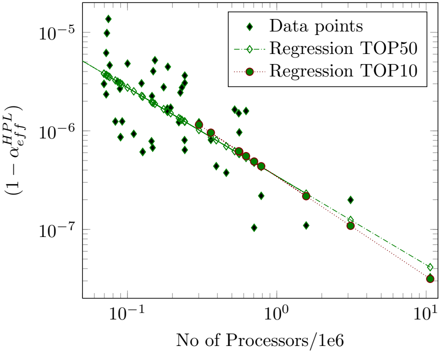
Correlation between accelerating, ranking and performance

Data copying between address spaces increase[10] payload performance by a factor 2...3, but it increases internal latency

[10] 10.1145/1815961.1816021

Do we need so many processors?





Correlation between processor number, $(1-\alpha)$ and ranking

A clear cutline between "racing" and "commodity" supercomputers

von Neumann and the modern technologies

- von Neumann explicitly warned (in his First Draft, section 6.3) that it would be unsound (sic) to use his classic paradigm to describe neuronal operation, given that the conduction (transfer) time is longer than the synaptic (processing) time.
- He also added that using much faster vacuum tubes (meaning processing elements) also vitiates his simplified paradigm, given that:
- In his simplified *classic paradigm* he omitted the transfer time (valid for that-time technology)
- However, *he did not provide another procedure* that could consider the case corresponding to today's technology and the case of neural computing
- We need to work out the way how transfer time can be considered; We need to introduce the *temporal behavior* into computing science

Generalizing the model

Actually, von Neumanns statement was that

if [timing relations of] vacuum_tubes

then Classic_Paradigm; //time-unaware

else Unsound;

Generalizing the model
Our claim is
if [*timing relations of*] *vacuum_tubes*
then Classic_Paradigm; //time-unaware
else Time − aware_Paradigm; // time aware

In this sense, we introduce time-aware computing paradigm.
We show that it describes both modern technological computing and biological computing.
J. Végh, Revising the classic computing paradigm and its technological implementations, in review in MDPI Informatics,
https://arxiv.org/pdf/2011.08455

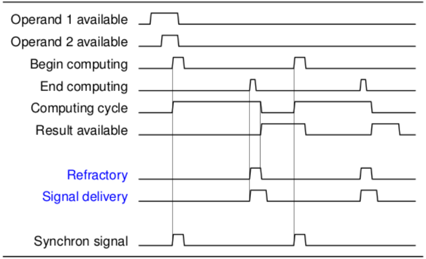


Fig. 1. Timing relations of Neumann's simplified (incomplete) timing model: the date transfer time neglected apart from data processing time

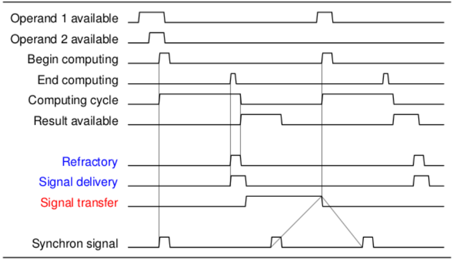
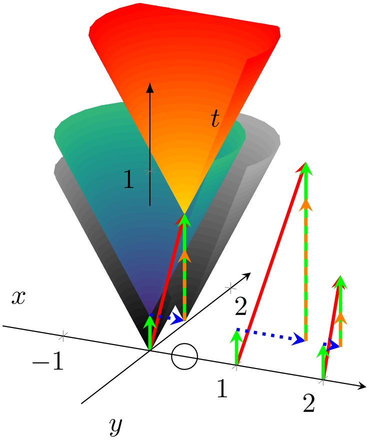


Fig. 2. Timing relations of Neumann's complete timing model, with data transfer time in chained operations

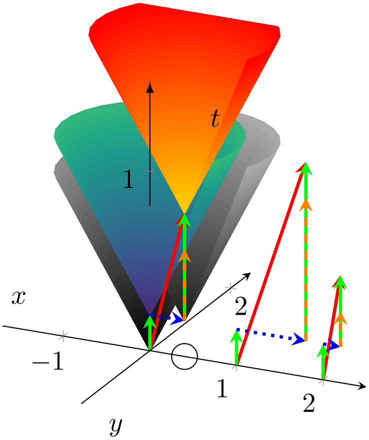
Computing model, without and with signal transfer. The compute and transfer operations block each other.[a]

[a] J. Végh: "A model for storing and processing information in technological and biological computing systems" Foundations of Computer Science, FCS4378



Introducing 4-D coordinates to describe computing

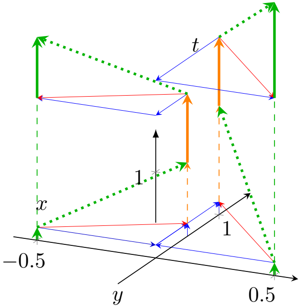
- Computing events have position and time
- Processing happens at the same place
- Transfer changes both positions an time
- Alignment constraint yields in "idle times"
- Observer's position matters



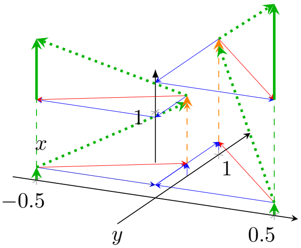
Introducing 4-D coordinates to describe computing

- Computing events have position and time
- Processing happens at the same place
- Transfer changes both positions an time
- Alignment constraint yields in "idle times"
- Observer's position matters
- Actually, Minkowski coordinates are used: interaction speed connects the coordinates

The apparent operation speed is not proportional to the physical operating speed of the memory

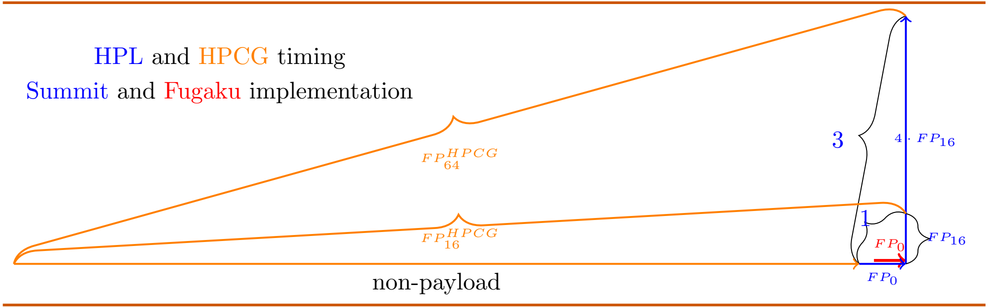


The temporal operation of a slow memory



The temporal operation of a fast memory

Do we need shorter operands?



Correlation between accelerating and ranking

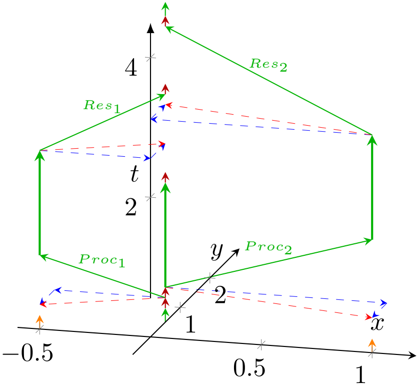
The shorter operand matters until transfer time starts to dominate
Only marginal improvement can be expected in huge systems form changing to half precision
FP_0: the "zero length" precision

AMD accelerator with different operand lengths

Type	AMD MI100	AMD MI50	AMD MI25
FP16 (4)	184.6 (16)	26.5 (4)	24.6 (4)
FP32 (2)	23.1 (2)	13.3 (2)	12.19 (2)
FP64 (1)	11.5 (1)	6.6 (1)	.768 (1/16)

The size of test data sometimes does not fit in the dedicated memory (32 GB and 16 GB) [11]

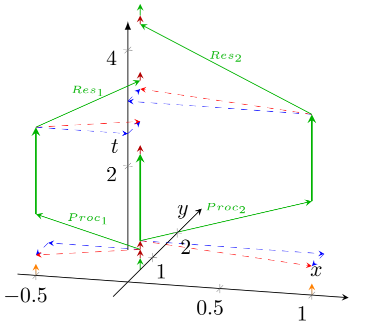
[11] https://www.amd.com/en/graphics/servers-radeon-instinct-mi



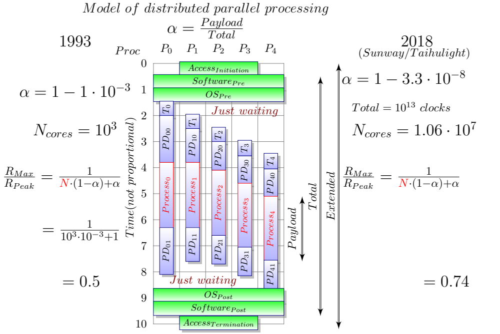
The temporal operation of distributed system

- The orchestrator starts computing
- The elements must be addressed
- The parameters must be sent out
- The results must be collected
- Parallelization introduces idle time
- The sequential activity is a bottleneck

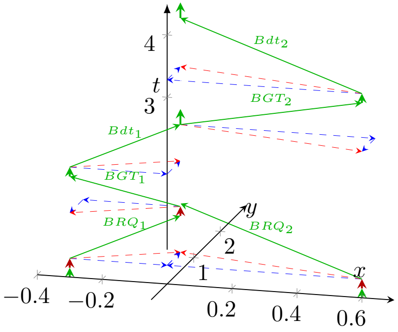
α describes the "goodness of parallelization" (non-payload/payload)



The temporal operation of distributed system



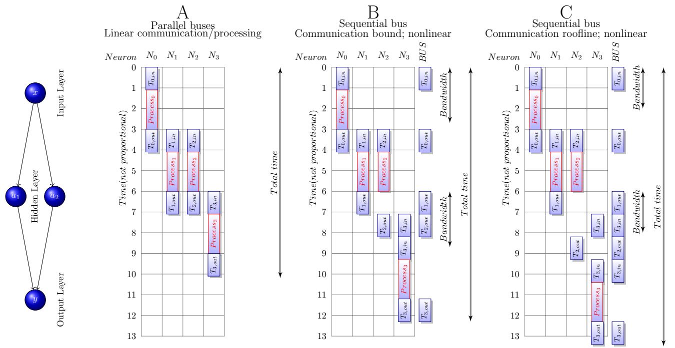
Temporal Operation Model



The temporal operation of distributed system

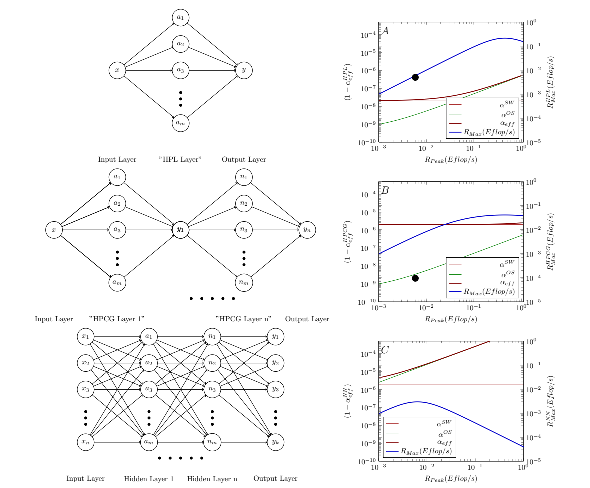
- A single medium is shared
- Bus arbitration needs a lot of time
- The contributors must wait for each other
- Most activity is spent with contenting
- The limiting factor is NOT the bus speed

Neural-type transfer on a sequential bus



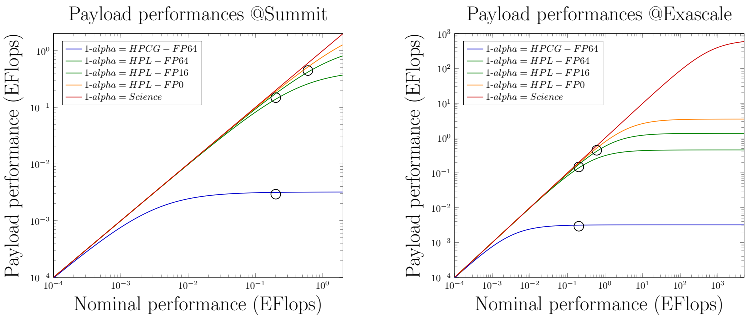
Neural communication. A (the biological implementation): parallel bus; B and C : the case of shared serial bus, before and after reaching the communication "roofline".

J. Végh, Which scaling rule applies to Artificial Neural Networks, Neural Comput & Applic. (2021) 10.1007/s00521-021-06456-y



Efficiency and performance of different workloads

- HPL benchmark has a minimum communication
- HPCG imitates real-life programs
- Increased communication decreases efficiency
- Inflexion and break-down exists
- ANN workload depends on M^2
- AI is the most disruptive workload (Cray)



Future of supercomputing, using classic paradigm and design principles.

Saturates towards high nominal performance

Workload selects diagram line (a "quantum system")

Both HW and SW brain simulation

Communication kills performance

The commonly used 1 msec 'grid time' is a clock period, 10^6 fold longer than the electronic clock period

SpiNNaker's α is ≈ 0.9; no way to work with billions of neurons

Temporal behavior of neuron

The "spatiotemporal" behavior in biology does not connect time and position as discussed here

The interaction speed is ≈ 10M fold higher than in electronic circuits;

Due to the mixed electrical/chemical charge transfer, the "non-Ohmic" behavior explains neuron's features ⇒ the real neural computing means adjusting temporal relations

Considering temporal behavior explains how information is stored and processed; Short and long term learning understood; matches anatomic evidence J. Végh and Á. J. Berki, On the Role of Information Transfer's Speed in Technological and Biological Computations, under review in SN Neuroinformatics, 10.20944/preprints202103.0415.v1

It can be pinpointed why learning and machine learning are different; explains awfully long training times

J. Végh and Á. J.Berki, Why learning and machine learning are different, Adv. Artificial Intelligence and Machine Learning, 1/2 (2021)9

Summary

- Computing, even the parallelized one, reached its limits
  (no financed grant or political will enables to exceed it)
- The need for new ideas/architectures/material is obvious
  However, no breakthrough idea
- The real issue is the nonlinear performance scaling;
  not noticeable at smaller nominal performances
- We cannot neglect time anymore in computing paradigm;
  we need a new (time-aware) paradigm
- The flown neglection leads to power and performance loss
- Introducing time to computing enables us to set up a general model for computing processes; including biological computing
- We need to reinvent computing, from the first principles