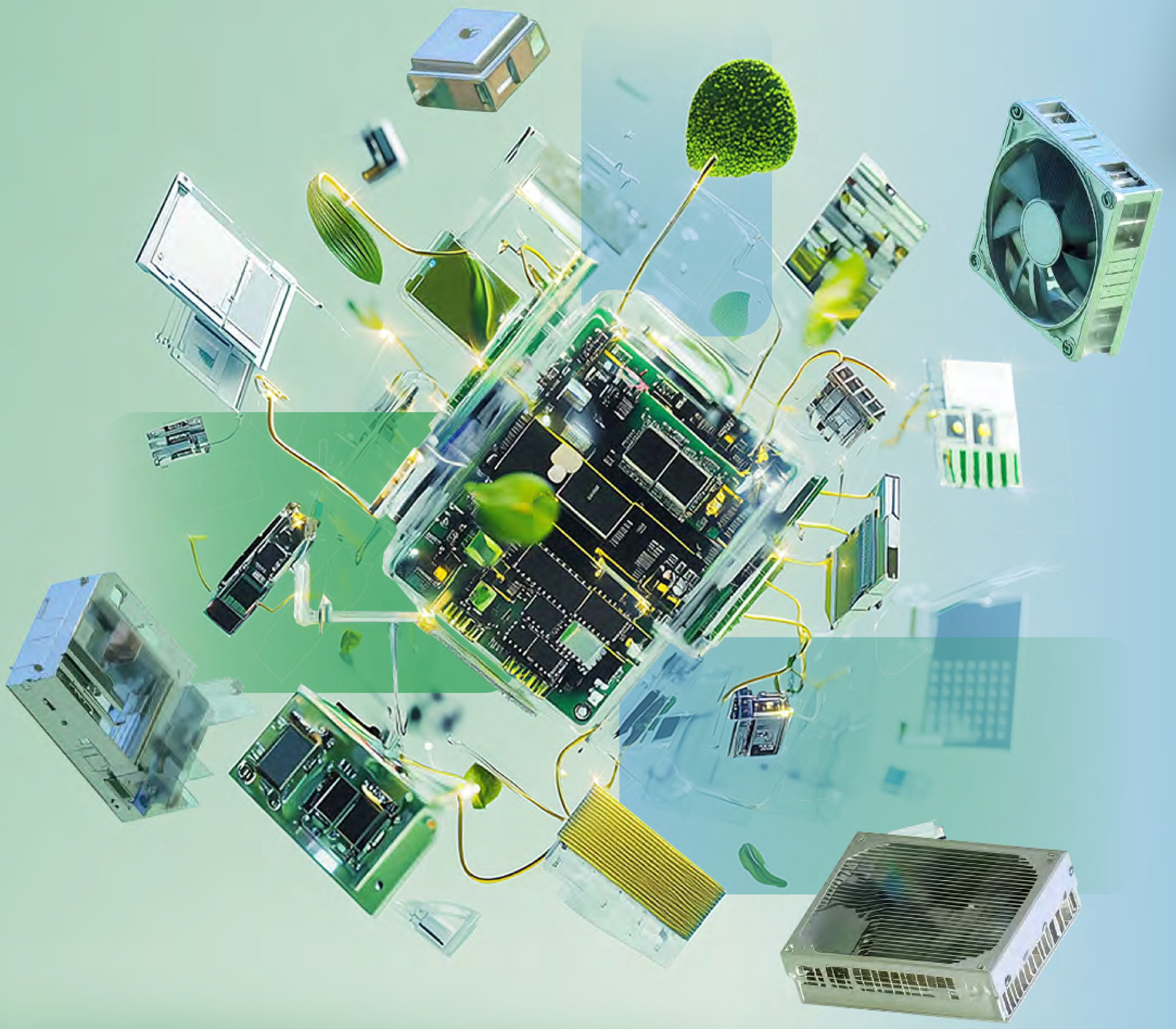
Computing

Authors
Irene Bonati (SURF), **Sagar Dolas** (SURF), **Amirreza Yousefzadeh** (University of Twente), **Guangzhi Tang** (Maastricht University), **Francesco Regazzoni** (University of Amsterdam), **Johan Mentink** (Radboud University)



1. Increasing need for heterogeneous and hybrid computing platforms
2. Trade-offs in accuracy with approximate computing
3. Growing emphasis on neuromorphic (brain-inspired) computing
4. Integrating photonics for accelerated information processing
5. Exploring biological systems for computing

SURF

Introduction

Society is increasingly relying on advanced computing capabilities to support the transition towards renewable energy systems, develop personalised medicine, utilise digital twins, and predict the impacts of climate change. This involves leveraging the state-of-the-art digital infrastructure ecosystem through collaborative, connected platforms, services, and products that enable analysis and virtual experimentation. Exascale computing platforms play a crucial role in processing information at scales previously unimaginable.

Within this landscape, computing is reaching a pivotal point where the combination of data, AI, cloud, and quantum-enabled supercomputing will play a more significant role in our ability to process information. This critical shift is driven by rapid advancements in AI, with a surge in demand for computational power that exceeds the limits of classical methods. Further, Moore's Law is slowing down, as silicon-based transistors approach physical limits. This drives the exploration of emerging technologies and paradigms, as well as their impact on computing for science, research and education.

At the same time, computing facilities are under pressure due to the large energy footprint required for data processing and infrastructure cooling. This is pushing the boundaries of traditional approaches and encouraging efforts towards more resilient and energy-efficient infrastructures.

Globally, there is a growing emphasis on advanced computing ecosystems as strategic assets for science, innovation, and growth,

along with ongoing advocacy for the required long-term investment. Computing, data and digital infrastructures are gaining more recognition at international forums, providing pathways for knowledge and sustainable societal development. The European Union, Asia, Africa, America, and other geopolitical regions have initiated public dialogues and launched projects to strengthen their computing ecosystems, recognising their impact on research, the economy, and society.

In the Netherlands, NWO (the Dutch research council), in collaboration with SURF, published a report highlighting the computational needs for accelerated scientific discovery. Alongside NWO, more researchers and research communities are emphasising the growing importance and demands of computing for scientific research across all disciplines.

Concurrently, the ICT Research Platform Netherlands has underscored the need for long-term strategic investments in the Netherlands, compared to international benchmarks, particularly in relation to Europe's mission towards digital sovereignty, knowledge, and economic competitiveness.

All these developments are co-evolving, influencing technology, policy, business models, and capital allocation, with implications for geopolitics, global talent competition, and economic and scientific impact. This chapter focuses on technological shifts in computing and their potential effects on research and education ecosystems.

Contributors
Patty Stabile (TU Eindhoven), **Matthias Möller** (TU Delft), **Manolis Sifalakis** (Innatera), **Alberto Bosio** (Ecole Centrale de Lyon), **Tom de Greef** (TU Eindhoven)

TREND #1

Increasing need for heterogeneous and hybrid computing platforms

Public Values

Autonomy	Independence | Freedom of choice | Privacy
Justice	Sustainability | Integrity | Accountability | Equity
Humanity	

Maturity WATCH PLAN **ACT**

Drivers

Engineering advances & computation; Digital transformation; Energy supply & demand;

The demands of modern computing applications have effectively challenged the era of classical approaches. Modern large-scale computing environments are increasingly becoming modular, integrating a wide range of specialisations such as hybrid quantum systems, cloud platforms, ASICs, GPUs, edge IoT devices, and storage, enabling the execution of complex tasks. The slowdown of Moore's Law, coupled with the growing demands of AI-driven design, necessitates specialised solutions for improved outcomes. Although the landscape remains fragmented, tools that facilitate orchestration across heterogeneous environments are becoming more widespread and valuable in mainstream scientific computing. We envision a lot of innovation in the software layers of the digital infrastructure stack for the coming decades.

SURF

SIGNALS

Hyperscale clouds design custom accelerators

AWS launches second-generation Trainium2 and Inferentia3 chips to reduce AI training costs (semianalysis.com)

Microsoft Azure introduces the Maia 100 AI accelerator and Athena CPU for its data centres (semianalysis.com)

Consumer Systems-on-a-chip integrate CPU + GPU + NPU

Apple's M3 Ultra features 32 CPU cores, 80 GPU cores, and an enhanced Neural Processor Engine (apple.com)

Qualcomm Snapdragon X Elite for Copilot+ PCs ships with a 45 TOPS NPU (qualcomm.com)

Converging software ecosystems

Intel oneAPI 2024 LTS unifies compilers across CPU, GPU and FPGA targets (intel.com)

Khronos releases SYCL Rev 9 with back-ends for CUDA, HIP and oneAPI (khronos.org)

"The classic computing approach, using program-flow simulations to model different functions, has become outdated, and the rise of AI and data-driven methods has accelerated the exposure of the cracks in the classic approach."

- **Manolis Sifalakis**, Innatera

Emergence of AI Driven design

AI-dedicated large-scale computing systems start to enter the market to improve efficiency and reduce cost compared to GPUs

- **Cerebras starts building six data centres with its wafer-scale AI hardware in North America and Europe from 2024 to 2025** (cerebras.ai)
- **Meta introduces next-generation AI infrastructure with custom-made AI hardware in 2024** (about.fb.com)
- **Tesla's wafer-scale Dojo AI hardware is in production to support large-scale training for autonomous driving in 2024** (tomshardware.com)

AI-dedicated hardware with emerging technologies is starting to transfer from academia to industry

- **SpiNNcloud introduces first commercial neuromorphic AI HPC in late 2024 using digital neuromorphic technology developed in the human brain project** (theregister.com)
- **In-memory computing technology starts to materialise into AI hardware products with start-up companies like Synthara and Axelera AI** (synthara.ai)
- **Innatera secures $21M in investments with plan to drive analogue neuromorphic AI to 1 billion devices by 2030** (unite.ai)

IMPACT

Education

Curricula are likely to increasingly embed specialised hardware concepts alongside conventional architectures. Students are therefore likely to engage further with a variety of accelerators dedicated to certain tasks in cloud lab environments. Lectures, labs, and projects will train students to deploy and optimise workloads on specialised accelerators. This shift is likely to drive academic programs to evolve beyond foundational programming courses towards more advanced parallel and heterogeneous computing paradigms. As a result, curricula will have to increasingly incorporate models such as CUDA, HIP, SYCL, or OpenCL.

Research

Scientific applications are likely to run more often on heterogeneous systems, such as “GPU for dense linear algebra, FPGA for bit-level genomics, or neural processing unit (NPU) for inference,” within a single cluster job. This shortens iteration cycles, enables larger parameter sweeps, and lowers energy budgets. Hardware-software co-design is set to become a fertile ground, inspiring resource-efficient algorithms and spurring rapid innovation in chip architectures. However, new scheduling policies and heterogeneous-aware middleware from campus HPC centres are needed as a result.

Operations

Operation teams might need to assess designs that combine different components while balancing power and cooling needs. Furthermore, facility layouts, monitoring tools, and support staff skills will need to adapt to efficiently manage mixed-architecture racks and the fast-evolving generations of accelerators.

TREND #2

Trade-offs in accuracy with approximate computing

Public Values

Autonomy		
Justice	Sustainability | Accountability | Equity	
Humanity	Safety	

Maturity WATCH PLAN **ACT**

Drivers
Energy supply & demand; Automation & AI; Engineering advances & Computation

There are computer applications that do not require high (computational) accuracy and precision, or work with information that already contains uncertainty. Think of applications like neural networks, signal processing, and localisation and mapping for which an approximate result is sufficient to meet both the requirements and design goals in the computer architecture.

Design goals can be traded, such as performance, power, and energy efficiency. Techniques available for approximate computing include quantisation, rounding, truncation, and reduction of the number of bits.

Approximate computing is also achieved at the software level with dedicated algorithms, and at the hardware level by using approximate hardware, or via a combination of both. Therefore, approximate computing is gaining more traction, leading to trade-offs in accuracy in exchange for other improvements in computational objectives.

SURF

SIGNALS

Reducing energy consumption using the approximate computing paradigm

Approximate computing-based accelerators to reduce energy consumption have been demonstrated through scientific research (ieeexplore.ieee.org) and projects (textarossa.eu)

Joint approximations across different subsystems have been explored to improve the energy efficiency of smart camera systems (ieeexplore.ieee.org)

Approximation used in combination with Dynamic Voltage Scaling

Dynamic error detection and correction is used for Dynamic Voltage Scaling in ARM Razor, that tunes the supply voltage by monitoring the error rate during operation (documentation-service.arm.com)

Aggressive voltage underscaling can be used in high-performance DNN accelerators while maintaining the required classification accuracy (ieeexplore.ieee.org)

AI Accelerators with approximate computing

IBM integrates approximate computing techniques in AI accelerators (research.ibm.com)

Approximate DRAM devices have been used to reduce the energy consumption and evaluation latency of DNN (arxiv.org)

"Approximate computing can significantly reduce the energy and computational cost of several applications, including machine learning, video processing, and data analytics."

- **Alberto Bosio**, University of Lyon

IMPACT

Education

Computer science is mostly based on determinism and exact computation. Students need to be trained to deal with approximated and not ‘exactly correct’ results. Concepts like error management and computation accuracy within known boundaries are likely to emerge both in theory and in practical sessions.

Research

The approximate computing paradigm is reasonably mature. Optimisation based on approximate computing might be beneficial for specific applications such as neural networks and could boost the performance of several research application domains, such as in the medical sciences.

Operations

Generally, the use of approximate software routines does not require an update of IT infrastructure. Larger benefits could come from the use of approximated hardware, but the changes of the infrastructure required must be contrasted with the resulting benefit. Such evaluation should be done individually on a case-by-case basis.

TREND #3

Growing emphasis on neuromorphic (brain-inspired) computing

Public Values

Autonomy		
Justice	Sustainability | Equity | Accountability | Integrity	
Humanity	Safety	

Maturity WATCH **PLAN** ACT

Drivers

Automation & AI; Engineering advances & computation; Energy supply & demand; Raw material scarcity; Clean water demand; Critical infrastructure

Neuromorphic computing is a computing paradigm inspired by the human brain. Just as the brain both processes and stores information, neuromorphic computers combine data storage and processing in one device, drastically reducing the energy needed for data transfer. This facilitates faster information processing and smoother handling of complex tasks like pattern recognition. It's less energy-intensive than digital systems.

Developments in neuromorphic computing are timely as modern data centres consume massive amounts of energy. Integrating neuromorphic processors in data centres could cut the energy needed for AI-processing up to 1000-fold.

SURF

SIGNALS

Potential to cut the required energy for AI-processing by 10 to 1000 times

The Netherlands bets on brain inspired computing for a greener future (computerweekly.com)

Kösters et al. (2023). **Benchmarking energy consumption and latency for neuromorphic computing in condensed matter and particle physics,** APL Machine Learning 1, pp. 16-101 (arxiv.org)

Intel builds worlds largest neuromorphic system to enable more sustainable AI (newsroom.intel.com)

Whitepaper Neuromorphic computing (2024) calls for the development of an open ecosystem and improved coordination in the Netherlands (ru.nl)

Startups commercialising neuromorphic compute

- axelera.ai
- innatera.com
- **Hoursec** (yesdelft.com)
- imchip.ai
- cimplic.com

Neuromorphic supercomputers

- spinncloud.com
- **DeepSouth** (westernsydney.edu.au)
- **Darwin Monkey** (globaltimes.cn)

IMPACT

Education

Several universities in the Netherlands are already expanding their educational programs to include neuromorphic computing and engineering and deepen the fundamental knowledge to educate the next generation of leaders. Furthermore, thanks to its energy efficiency, neuromorphic computing has the potential to make AI-based innovations, such as personalised teaching bots, locally available in the classroom.

Research

In 2024, neuromorphic supercomputers made their debut in HPC (Spinnaker 2 (Germany), Intel's Halapoint (US), DeepSouth (Australia), Darwin Monkey (China)), thus reaching over one billion artificial neurons. The ability to understand which applications benefit most from which neuromorphic hardware is still largely open, with calls for extensive benchmarking. Moreover, current neuromorphic systems operate still far above the energy efficiency of the human brain, calling for foundational discoveries to compute as efficiently and functional as the human brain.

TREND #4

Integrating photonics for accelerated information processing

Public Values

Autonomy	Independence
Justice	Sustainibility | Equity
Humanity	Safety | Well-being

Maturity WATCH **PLAN** ACT

Drivers

Automation & AI; Engineering advances & computation; Energy supply & demand; clean water demand; critical infrastructure

Photonics has already played a pivotal role in the internet cables that carry data globally and serve as a fundamental technology for data transmission within data centres.

Recently, photonics has attracted significant attention for computing applications. The same principles that benefit communication – low latency, energy efficiency, and the capacity to process data on the fly – are now being explored to overcome the limitations of electronic computing (such as energy consumption, latency, and bandwidth). This will benefit applications such as in the acceleration of neural networks, data centre workloads, and edge AI.

SURF

SIGNALS

Integrated use of photonics in the data centres

Integrated photonics is gradually making its way into the realm of supercomputing (marvell.com)

Running large-scale matrix multiplication with an energy advantage of O(1000x)

In contrast to electronics, photonic systems can naturally exploit parallelism, making them well-suited for real-time processing at scale (arxiv.org)

LUMAI offers a promising advancement in AI acceleration (lumai.ai)

Photonics in the knowledge and innovation agenda for key enabling technologies

(kia-st.nl)

"The tremendous parallelism offered by photonics, with each beam in space representing individual computations, may illuminate the future of computing."

- **Patty Stabile**, TU Eindhoven

IMPACT

Education

If optical computing continues to mature as a foundational technology for data processing and communication, education will inevitably need to evolve accordingly. The traditional focus on electronic-based computation in engineering curricula will no longer suffice to prepare students for the hybrid electronic and photonic systems of the future. At institutions such as TU Eindhoven, where an elective course on optical computing has already been introduced, this evolution has commenced. While still optional, such courses represent an early integration of photonics into the broader computing curriculum.

Research

The integration of optical computing into digital infrastructure promises significant advancements in research capabilities, enhancing the quality of infrastructure, computational tools, and processing power. Optical systems might enable faster, more energy-efficient data centres and clusters, which are vital for data-intensive fields such as climate science, genomics, particle physics, and AI. A hybrid electronic-photonic approach enhances simulation performance and energy efficiency, allowing researchers to address difficult-to-tackle challenges, such as high-resolution brain simulations or chemical modelling.

Operations

Photonic integration also enhances data centre operations by promoting sustainability. Optical components consume less energy, produce less heat, and reduce cooling needs, leading to lower costs and higher workload density. In addition, the decrease in energy requirements can lead to less stress for local or regional power networks.

TREND #5

Exploring biological systems for computing

Public Values

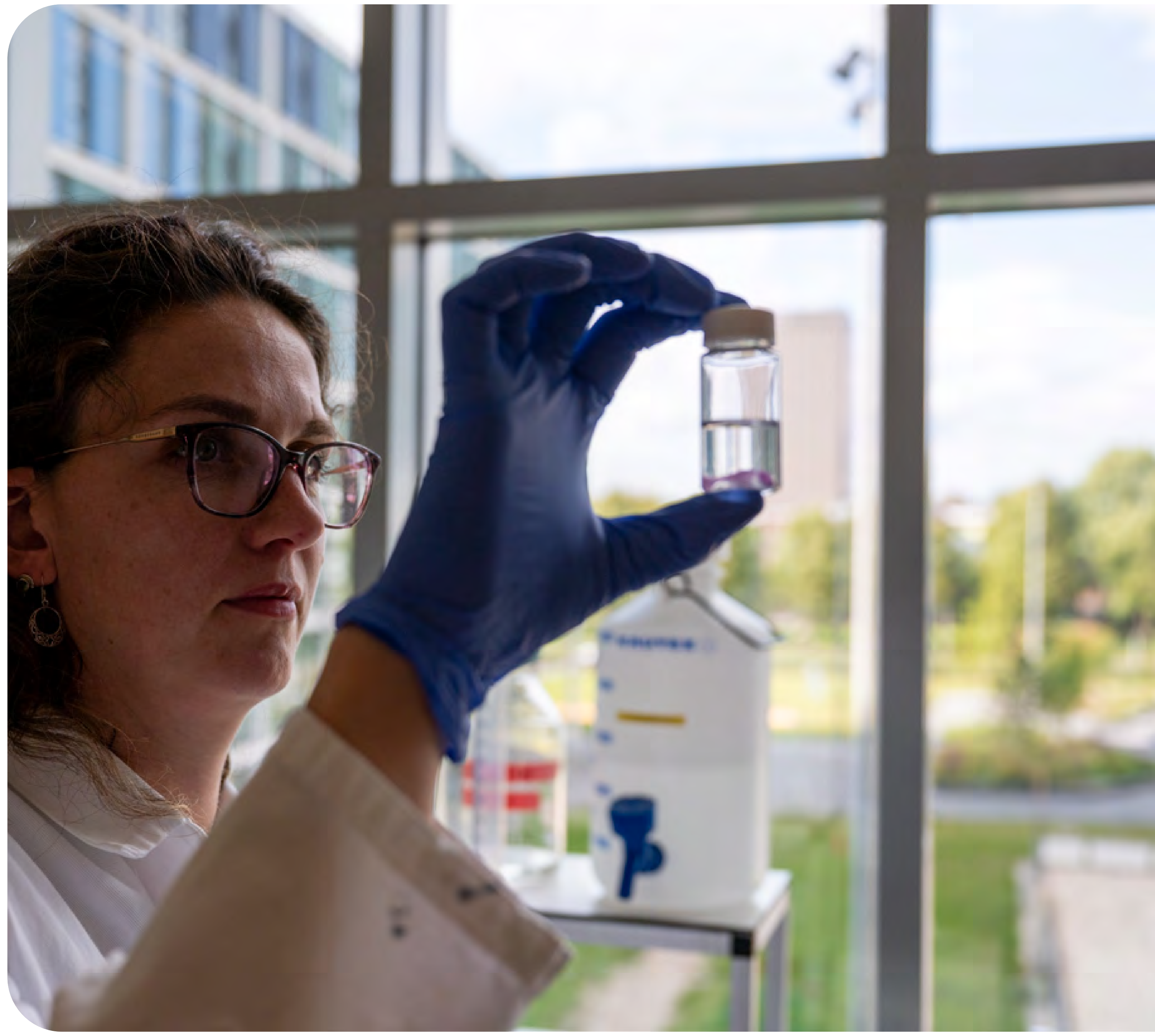
Autonomy	Privacy
Justice	Integrity | Sustainability | Equality
Humanity	Well-being | Safety

Maturity WATCH PLAN ACT

Drivers

Engineering advances & computation; biotechnology; Automation & AI; Energy supply & demand

Living matter also performs computation. Recent years have featured a rise in the exploration of the inherent information processing capabilities of biological material as a foundation for computing systems. Biocomputing is the use of biological components, such as DNA, proteins, and cells, to perform calculations, similar to a traditional computer, but on a molecular scale. This technique offers an entirely new way of encoding and processing information, yielding some improved capabilities compared to traditional architectures, such as massive parallelism and energy efficiency. Although still an experimental field with technical challenges to be overcome (e.g., speed, scalability), research related to biocomputing is expected to continue, with this approach getting closer to real-world applications, especially in healthcare, in the next decades.



SIGNALS

Early commercialisation by companies

The company FinalSpark is offering cloud-accessible computing services using real organoids through a platform (Neuroplatform) (finalspark.com)

The start-up Cortical Labs launched a biological computer made of human brain cells (abc.net.au)

"There are promising applications of biomolecular computing in healthcare, for example; Since biomolecular computers work with molecular inputs, they could easily measure biomarker profiles and make decisions about whether there is a certain disease state or not. [...] The main bottleneck is that currently speed is very limited, even if computing operations can be massively parallelised"

- **Tom de Greef**, TU Eindhoven

R&D milestones modifying biological processes to perform computing-like routines

Synthetic biology used to build genetic circuits, equivalent to logic circuits in conventional computers (news.mit.edu)

Researchers developed a software package that facilitates the design of digital logic circuits in molecular computing (dl.acm.org)

Researchers genetically modified a strain of E. coli for computation (nature.com)

Researchers outlined a vision for biocomputers powered by human brain cells and presented a roadmap for organoid intelligence (frontiersin.org)

A biocomputing system made of living human brain cells performed speech recognition starting from audio clips (nature.com)

Researchers achieved repeated data operations (storing, retrieving, computing, erasing, and rewriting) using DNA (nature.com)

Researchers designed the first programmable DNA computer (nature.com)

IMPACT

Education

Cross-disciplinary talent skilled in both computing and biology is likely to increase in demand. This necessitates the development of new or extension of existing curricula for students in scientific disciplines benefitting from biocomputing to acquire this hybrid skillset. Computers using biological circuits may also emerge as hands-on teaching tools to illustrate the principles governing biocomputing.

Research

Once the current efficiency and scalability issues are overcome, researchers might be able to access local and/or national IT infrastructure working following the principles of biocomputing. This will be beneficial to a variety of research use cases in different scientific areas, such as drug discovery, disease diagnosis, and ecology.

Operations

The shift towards biocomputing might necessitate IT infrastructure upgrades and require (research support) organisations to re-evaluate operational processes to handle biocomputers. This requires staff skilled in working with biological systems and/or possible reliance on external parties/organisations specialised in working with biocomputers.

More info about Computing Technologies? Visit surf.nl

SURF Utrecht
Hoog Overborch Office Building (Hoog Catharijne)
Moreelsepark 48
3511 EP Utrecht
+31 88 787 30 00

SURF Amsterdam
Science Park 140
1098 XG Amsterdam
+31 88 787 30 00

futuring@surf.nl
www.surf.nl/en