Photonic Computing for Massively Parallel AI

A White Paper

March 2021
Photonic Computing for Massively Parallel AI

A White Paper

March 2021

LightOn SAS
Headquarters: 2, rue de la Bourse F-75002 Paris
Offices: 5, Impasse Reille F-75014 Paris
E-mail: contact@lighton.ai

Whilst reasonable effort has been made to ensure that the content of this White Paper was correct at publication date, neither LightOn SAS nor any person engaged or employed by LightOn SAS accepts any liability for any errors, omissions or other inaccuracies. Readers should independently verify any facts as no liability can be accepted in this regard. This White Paper may include predictions or estimates that might be considered forward-looking. While these forward-looking statements reflect our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ. Readers are cautioned not to place undue reliance on these forward-looking statements, which reflect our opinions only as of the date of publication. We are not obligating ourselves to revise or publicly release the results of any revision to these forward-looking statements in light of new information or future events. All product and company names and logos appearing on this White Paper are trademarks, service marks or trading names of their respective owners, including LightOn SAS.
<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>There’s light beyond pure silicon</td>
<td>4</td>
</tr>
<tr>
<td>02</td>
<td>Million pixel baby: LightOn’s Massively Parallel AI accelerator in action</td>
<td>6</td>
</tr>
<tr>
<td>03</td>
<td>Because it’s time: Sketches as first class citizens</td>
<td>7</td>
</tr>
<tr>
<td>04</td>
<td>A data scientist shouldn’t have to wait</td>
<td>9</td>
</tr>
<tr>
<td>05</td>
<td>LightOn Cloud: Play and invent the stuff it was never intended for</td>
<td>15</td>
</tr>
<tr>
<td>06</td>
<td>So it begins</td>
<td>18</td>
</tr>
<tr>
<td>07</td>
<td>Who we are</td>
<td>20</td>
</tr>
</tbody>
</table>
There’s light beyond pure silicon

“The world is rapidly running out of computing capacity.”
Satya Nadella, Microsoft’s CEO, January 2018

“It means that at some point we’re going to hit the wall. In many ways we already have [...] and we really need to look at how we get most out of the compute we have. This is the world we are going into.”
Jerome Presenti, Facebook’s Head of AI, Wired, December 2019

Who could have imagined these striking statements only a decade ago? For the past 60 years, the world has relied on the ingenuity of the electronics industry to produce exponentially more powerful chips without increasing costs, i.e., Moore’s law. Matching the computing demand was only a matter of producing more chips or waiting a couple of years for the next generation. Today, a conjunction of a much faster increase in demand and a struggle for supply is causing a gridlock.

The supply: in recent years, the silicon industry has been failing to keep up with the promise of Moore's law. It now has to resort to a myriad of workarounds, such as increased parallelism, reduced precision or rapid development of nascent technologies, such as compute-in-memory or 3D stacking, to make this happen. At the same time, capital expenditures for 10 nm and below technologies drive up production costs while providing only limited capacity. General purpose CPUs (Central Processing Units) are not sufficient anymore. As a result, specialized hardware dedicated to the primitives of AI (Artificial Intelligence) has fueled a “Cambrian Explosion” of AI chips, with more than a hundred projects worldwide. Interestingly, these specialized chips have sprung up in a wide variety of contexts. While some were born from projects started at incumbent chip manufacturers, some of the most prominent ones originate from startups (Habana Labs—just acquired by Intel—, Cerebras, Graphcore, etc). More surprisingly, some of the new AI chip designs started at large data-centric corporations such as Google (TPU), or Alibaba (Hanguang 8000), where the demand is strongly felt.

The demand: unlike previous times, increase in compute demand outpaces the technology. Rich data sensors have become ubiquitous, and business needs require the rapid analysis of these large-dimensional streams, billion-scale users/products recommender systems, etc. Examples abound, but here are two that are a sign of the times: on the last 11/11 Singles’ day (2019), Alibaba sold more than $38B in merchandise through its ability to scale personalized recommendations to more than 700M smartphones that one day (Figure 2). In another instance, Google is said to have avoided the construction of a dozen additional data centers thanks to the optimized processing power offered by the TPU1. For AI to reach its potential of an increased $3.5-5.8T in annual business revenue2, a massive additional compute capacity is simply essential. The technology pull is illustrated in a recent study by Business Wire3 that estimates the AI accelerator chip market growth to be at CAGR of 35% to $59B by 2025.

Most new computing architectures are based on digital logic and silicon technology. Their progress is still bound by fundamental limitations that include both an increase in manufacturing costs, as well as memory issues of traditional Von-Neumann architectures. As a result, a handful of technological platforms are trying to solve these challenges. Quantum computing is one such new technology, yet it is still years away from the scalability and flexibility that businesses currently require.

For computation technology, a “beyond pure silicon” scenario requires components and designs that have a high readiness level and, most importantly, can leverage the 60+ year

Figure 2: With more than $38Bn merchandises sold on Alibaba’s Single’s day (Nov 11), E-commerce comes out as a winner in the new economy. At the core of these staggering numbers is the capacity to make personalized product recommendations at scale for millions of users. Source: Techcrunch, https://techcrunch.com/2018/11/11/alibaba-singles-day-2018-31b/

Figure 3: According to an analysis from the Omdia / OVUM Report4, “January 2020 AI Reality chart tracking the analog and photonic AI hardware accelerator race”. LightOn has the most mature technology in this category.

2. McKinsey Global Institute, “Notes from the AI frontier: Applications and value of deep learning”, April 2018
investments in electronics. Today, the only such technology is photonics. Light can have extremely weak interactions with matter: moving information around produces virtually no heat, and therefore enjoys extremely low power consumption. Since the late 1990s and the rise of the Internet, photonics has surfed on Moore's law, as exemplified in the development of Micro-Electro-Mechanical Devices (MEMS) or silicon for imaging, and is present in billions of smartphone cameras. In order to accelerate computing capabilities and enable a continuing growth in terms of performance beyond Moore's law, it is time to use photonics beyond communication and sensing. While historically photonics has changed information transfer, it can now be an additional enabler of computing progress.

LightOn positions itself as the leader in the transition away from pure silicon computing, enabling continued growth in computation performance.

At LightOn, we have built the first optical AI chips running in data centers that are capable of performing Large Scale Machine Learning computations. In the last two years, some of these Optical Processing Unit (OPU) prototypes have addressed dozens of different use cases in Machine Learning. Early OPU users include external beta-users from academia and industry in Europe, America, and Asia. LightOn's current OPU (“Aurora”) is a specialized co-processor whose photonic core (“Nitro”) performs one class of computing operations in an extremely efficient manner. In terms of speed: LightOn’s photonic cores can process data up to two to three orders of magnitude faster than standard processors on this operation.

Where’s the magic? In a nutshell, LightOn’s technology harnesses a physical phenomenon called “multiple light scattering”. It does so using mature technologies that have large and established supply chains. As is the case in electronics, our engineers made it so that typical users do not have to understand the underlying physics of the photonic core. In practice, data scientists can use the OPU with a single line of Python code, thanks to LightOn's proprietary library. Much like any other specialized hardware such as Graphics Processing Units (GPUs), LightOn OPUs connect to a computer through one of the motherboard’s PCIe slots.

The simplicity of the software layer stack and the maturity of the hardware and its attendant supply chain are both important. However, the powerhouse behind LightOn OPUs revolves around how large-scale data can be seamlessly handled, used, and processed through their transformation into nearly equally informative, smaller sketches.
The history of computing has evolved with the increased complexity of data structures over the past 60 years. CPUs are ideal processors for scalar computations. GPUs are optimized for vector computations and have indeed been a key enabler technology for Deep Learning. Today, we introduce LightOn Optical Processing Units (OPUs), optimized for the new world of high-dimensional data. High-dimensional data is not only expensive to work with, it is also unwieldy and difficult to process. As a consequence, and paraphrasing Rutherford D. Rogers/John Naisbitt, one gets drowned in information but starved for knowledge.

In order to break this "curse of dimensionality" and allow for rapid use of data, researchers have traditionally sought ways to reduce its dimensionality without compromising the underlying information. Techniques such as Principle Component Analysis (PCA), projection pursuit, or trainable autoencoders are part of the common toolkit leveraged by data scientists. Ironically, these techniques fail to scale up to truly high-dimensional data, for they become too expensive to compute. Mathematically speaking, LightOn's first series of photonic cores computes sketches of large scale data through an operation called "Random Projection". This transform amounts to multiplying an input data vector by a fixed matrix full of random numbers. The output size of the result is a user-controlled parameter that drives the compression ratio: the smaller the size, the higher the compression. There is a whole range of mild compression rates where the information can be fully preserved. This is known since the compressed sensing theory emerged in the mid-2000s and which itself is well grounded in Information Theory and Mathematics ("Johnson-Lindenstrauss lemma").

A LightOn OPU can therefore be seen as a universal compression engine. In other words, its photonic core computes compact sketches of high dimensional data with no a-priori knowledge of said data all the while retaining most informational content, making all subsequent processing easier. At stronger compression rates some details may be lost, while still ensuring enough information for clustering or classification. Remarkably, LightOn's photonic core can also work in dimensionality expansion, approximating a well-defined kernel with so-called random embeddings.

![Figure 4: Some examples of computing architectures that incorporate LightOn OPUs in the data processing pipeline.](image-url)
The use of these universal sketches for practical algorithms is **central for any task that deals with high-dimensional data.** By making data smaller, LightOn OPUs unlock all the regular Machine Learning algorithms from Unsupervised Machine Learning (Singular Value Decomposition, Nearest Neighbors, Matrix Preconditioning, Randomized Numerical Linear Algebra) as well as those used in time series modeling (Anomaly Detection, Reservoir Computing, Echo-State Networks, etc.), and those of traditional Supervised Machine Learning (k-NN, Transfer Learning, Reinforcement Learning).

Furthermore, since Deep Neural Networks may need to handle a large amount of high dimensional data, their training can be sped up by LightOn’s OPU through the replacement of the unwieldy back-propagation operation with an embarrassingly parallel operation. Because the OPU transform enables algorithms to work natively on a separate representation, random projections can also have applications in differential privacy.

While the vast majority of AI chips focus on the markets of Supervised Learning where training and inference are key, aside from the hardware technology, LightOn also positions itself on the often overlooked areas of Online and Unsupervised Machine Learning techniques that some see as the future of Artificial Intelligence:

“I also think the only path to developing really powerful AI would be to use this unstructured information. It’s also called unsupervised learning.”

Demi Hassabis, DeepMind CEO, Wired, January 2015

---

**In this new world where businesses require fast actionable items from rich data, LightOn’s hardware makes these sketches actionable as they become as natural as floats or int16 are for CPUs.**
Data scientists are a unique breed in how they try to marry data, models and reality. In doing so, what they really need are tools for the rapid prototyping of their ideas and models. Abe Stanway said it better than we ever could:

“All hard problems are just slow feedback loops in disguise.”

How can this slow feedback loop be avoided in every new investigation? How can one “do machine learning efficiently”? Radek Osmulski explicits his 10-second rule for Machine learners:

“The solution might sound ridiculous but it works. Do not ever allow calculations to exceed 10 seconds while you work on a problem.”

Benchmarks: Nitro-boost your GPU on everyday use-cases!

With inputs and outputs each in the one million range, LightOn’s current “Nitro” photonic core performs the equivalent of $2 \times 10^{12}$ hybrid-precision MAC (multiply-accumulate) operations at every clock cycle. Since the OPU runs at 1.5 kHz (Aurora 1.5), its “Nitro” photonic core yields a stunning performance of $3 \times 10^{15}$ MAC/s with only 30 Watts of power or $10^{14}$ hybrid-precision MAC/J. This performance is two to three orders of magnitude better than the best pure silicon AI chips such as Google’s TPU. In the next examples, we show how this performance translates into real gains for complete Machine Learning pipelines.

This first class of benchmarks presented here is only the tip of the iceberg. Indeed, since the OPU can handle very large data, users can venture into processing massive datasets that only a happy few, blessed with large computing resources, have previously been able to explore.

And this is a game changer.

---

6. It should be emphasized that these comparisons should be taken carefully, as the OPU photonic core operates in a “non-Von Neumann” regime, i.e. where the weights of the matrix - the multipliers - can neither be programmed nor tuned. Instead, one may see the photonic core of the OPU as a very large Read-Only Memory storing $10^{12}$ random coefficients that can be accessed literally “at the speed of light” and in an energy-passive way. Speed bottlenecks and power consumption arise as a result of communication; formatting; digital-to-analog and analog-to-digital conversion steps, as well as the need to power the light source.
Transfer Learning on images

Transfer Learning solves the problem of training state-of-the-art Neural Network architectures for smaller datasets. In Transfer Learning you start with an already trained network and only adapt a small subset of its parameters to the new data at hand.

Let’s take the example of a standard Deep Neural Network architecture such as VGG16. This network has first been pre-trained on the ImageNet dataset, composed of millions of labelled images with generic content. We want to adapt this architecture to another, much smaller dataset (STL10, here with “only” 13,000 images). To do this, we keep the first pre-trained convolutional layers, and only update the dense layers that link the features (25,000 per image) to the new classes of interest. For full reproducibility, this example is detailed on our blog with accessible code on our public GitHub repository. The result? One can train the network with the OPU in less than 30 seconds instead of nearly 4 minutes with a GPU only, a 8x speedup and a total energy cost divided by more than 20! It should be stressed that this result does not come at the expense of a loss of classification accuracy, even though LightOn OPU works in the analog domain.

Similar results are obtained on other image datasets, or other network architectures. For instance, on the ResNet152 architecture, we observe a 10x speedup in training time and 12x energy savings for the “skin cancer” dataset, and a 10x speedup and 15x energy saving for the “flowers” dataset.

For the problem of action recognition in videos, with just one OPU and one GPU, it is possible to obtain a baseline result in a couple of minutes, compared to many hours on potentially multiple GPUs—as reported in similar studies found in the literature. For similar accuracy, when comparing with standard methods with a well chosen set of hyper-parameters, the OPU has been shown to provide a speedup above 9x. When those hyper-parameters are not found right away, getting the right combination can take to data scientists many hours. This is a good case where obtaining a baseline in minutes allows for faster and more creative interaction with the model and the dataset.
Approximating Kernel Ridge Regression for classification tasks

In this example, we combined the power of CNNs and of the random features to increase the accuracy of a simple CNN by replacing the last dense layer of common pre-trained networks by a kernel ridge regression classifier, where the random projection is performed using the OPU. And trying to solve Ridge Regression problems with hundreds thousands of features without hitting the GPU memory limit is not an easy task!

Figure 7: Performance gains on Kernel ridge approximation for classification tasks. Dataset qm7 (quantum chemistry), high energy physics, and others. The OPU is compared to an NVIDIA P100 GPU. GPU RAM limit was hit at 32GB. Results acquired extrapolating to 1M features.

Dimensionality Reduction

Supervised Random Projections

As high-dimensional datasets become more common, dimensionality reduction becomes greatly important. In addition to easing memory and computing constraints, it enables visualization and often improves modelization. However, mainstream techniques, as PCA, are not ideal for supervised learning, where techniques as Supervised Random Projections (SRP) perform much better. Using an OPU, SRP become even more useful, and the whole pipeline is greatly accelerated, as can be seen in the fig 8, when we reproduced some literature results.

Figure 8: 1.2x to 4x speedup at label dimensionality 100k to 700k. Dataset: dimensionality reduction benchmark datasets, Library: LightOnML.

Randomised SVD for Recommender Systems

Recommender systems are becoming prevalent in the e-commerce business. In this example we tried to offer a quick and easy baseline for large-scale recommender systems using our OPU. Initially, we tried to apply the Singular Value Decomposition algorithm (SVD) on the MovieLens-20m dataset but the task failed due to the high memory requirement of the algorithm. Using randomised SVD instead (reducing data dimensionality through random projections before performing SVD) reduces the memory requirements and decreases the computational complexity.

Figure 9: Reduced memory requirements and computational complexity using randomised SVD that can be handled by the OPU. Database MovieLens-20m: 27000 movies, 138000 users, with 0.5% non-zero entries. Library: LightOnML.

Change point detection

Change detection in molecular dynamics

Our engineer Amélie has been investigating molecular dynamics in Computational Chemistry, used in drug design and new material discovery. Adapting the NEWMA\(^9\) change-point detection algorithm to simulation data from SARS-COV2 Molecular Dynamics trajectories, she was able to efficiently detect conformational changes, with a method that can easily scale to systems of very large numbers of atoms such as viruses and proteins, or to computations including solvent effects. And the results show that OPU enables analysis of samples containing a very large number of atoms due to overcoming the memory bottleneck of traditional architectures.


Figure 10: OPUs can be used for efficient and fast conformational changes detection in Molecular Dynamics simulations, even on very large systems overcoming the memory bottleneck of traditional architectures. These simulations, involving some of the largest High Performance Computing (HPC) resources, are nowadays a key technology in drug design.

Another change detection application where LightOn OPUs produce great results is change detection in time evolving graphs, with applications in community detection, fraud detection, biology, etc.

In all those applications it is of great importance to be able to detect real-time changes without the need to store the whole history of the graph.

Figure 11: 15x faster than FastFood on CPU at 50k atoms! For 700k + atoms, NEWMA RP on OPU is expected to be 30x faster than NEWMA FF on CPU. Library: LightOnML, Dataset: Molecular Dynamics simulations (HPC, Anton), Nitro photonic core, Aurora 1.5 (LightOn Cloud)

Figure 12: Using OPUs to detect real-time changes in time-evolving graphs requires less memory thus facilitating the analysis of very large datasets. Library: LightOnML, dataset: Facebook graph datasets, Nitro photonic core, Aurora 1.5 (LightOn Cloud)
Reinforcement Learning

Model Free Episodic Control

Our intern Martin investigated Reinforcement Learning on how to play a video game. Using the OPU to quickly search into past actions-rewards, he was able to successfully train a system to play classic arcade games such as PacMan or Space Invaders—with absolutely no rules given. At the end of the internship, Martin’s algorithm could clear the first level of Space Invaders thanks to computations performed on a LightOn OPU.

But why? Games provide toy versions of complex real-life problems and Reinforcement Learning (RL) is a powerful way to solve them. By this example we show how RL algorithms can be more lightweight using looking at the model-free episodic control algorithm. While Model-free episodic control is not the panacea of RL, it is a good starting point to understand the success of some methods that has been shown to outperform deep RL algorithms in some cases.

Neural Network training

Direct Feedback Alignment

While backpropagation has long been the standard training method for neural networks, new synthetic gradient methods—where the error gradient is only roughly approximated—gain interest. These methods represent better how biological brains learn, but also open new computational possibilities. Even so, they fail to scale past simple tasks like MNIST or CIFAR-10.

In this example we focused on Direct Feedback Alignment (DFA) which is one such method. The result? We were able to train successfully several modern neural architectures using DFA and a LightOn OPU, such as for Graph Neural Networks. Faster training or the training of very large language models that require data center scale resources will benefit from an all-optical device such as LightOn’s OPU.

Figure 13: Using the OPU for Reinforcement Learning. Library: LightOnML, dataset: Atari Games, Nitro photonic core, Aurora 1.5 (LightOn Cloud)

Figure 14: Optical training demonstrated on MNIST. Dataset: widely applicable (RecSys, graphs, NLP, etc.) Next-generation photonic core Lazuli 1.5 (Next-generation photonic core Lazuli 1.5 - available on LightOn Cloud for alpha users)

For sentiment analysis using Natural Language processing, our engineer François worked only a few days on word/sentence embeddings with the OPU and applied it to a new dataset. All the details are given in our blog post related to this experiment.

Key takeaways: it is very easy to obtain satisfying results quickly using the OPUs, without need for training, and for any language (as long as word embeddings are available).

The process of creating, as well as updating, models and their implementation requires the full attention of data scientists. Any delay, stopping them from being inventive, is bound to have an overly detrimental effect on their output and their satisfaction.

LightOn’s technology allows data scientists to test and iterate rapidly on new problems. In the aforementioned use-cases our engineers have developed models from the ground up in a matter of days from beginning to end, in domains were initially they were not proficient. Nevertheless, in less than a week, they obtained results that are close to the state of the art, or difficult to have without multiple GPUs. If you are new to a field, you can create a robust baseline rapidly and efficiently with LightOn’s technology.

Everything else may be a luxury.

8. https://www.lighton.ai/blog/
The large majority of Machine Learning Engineers and Data Scientists are accustomed to performing computations in the cloud through different services such as Microsoft Azure, Amazon Web Services (AWS) or Google Cloud. In order to showcase our technology to early adopters and decision makers, LightOn maintains a series of OPU prototypes that are remotely accessible 24/7 from anywhere in the world. This service is called LightOn Cloud.

On the hardware side, LightOn Cloud features a series of servers with the latest-generation LightOn OPUs, a high-end V100 Nvidia GPU, and Intel CPU. Users can easily benchmark LightOn OPU performance on their own data and code against state-of-the-art chips. More importantly, users’ inventiveness is unleashed as they can now design new models that rely on the combined strength and efficiency of each of these processing units (CPU, GPU, OPU).

On the software side of the LightOn Cloud, much effort has been undertaken to enable seamless integration of the OPU into the digital pipelines commonly used by Machine Learning experts. Hence, our OPUs are easy to use through our Python Library (LightOnML library). Access is possible through SSH login or a web-based Jupyter notebook. To provide this efficient service, LightOn Cloud benefits from housing partnerships with two of the largest cloud providers in Europe, OVH Cloud and Scaleway.

A typical use scenario and our value proposition

Laura is a Machine Learning engineer working at an e-commerce company. She is developing recommendation engines as the company catalog is constantly evolving. Laura accesses LightOn Cloud where she uploads a large dataset and her own code. She then begins to use the OPU through its Python API. During the day, Laura uses a combination of CPU, GPU, and OPU to speed up the training of efficient models, prototyping new ideas. Transfer Learning is a key technique to generate new Machine Learning pipelines from pre-trained models. Because the OPU compresses data, the large data she uploaded earlier on LightOn Cloud has now become...
We are hugely excited about the potential of Optical Processing Units (OPUs) to change the way we think and implement Machine Learning models. [...] OPUs offer a truly promising solution, by offering randomized computations with low power consumption...

Nicolas Keriven, CNRS researcher with the CICS team (Communication and Information in Complex Systems) in the GIPSA laboratory, Grenoble, France.

At some point, the uploading of her data is going to take too long because it is becoming too large. Laura’s company can then decide that they ought to have their own LightOn OPU on-premise (LightOn Appliance) so that the rapid prototyping can be done fully in-house on potentially much larger, constantly evolving datasets.

Laura’s case is just one example of the value brought by LightOn’s technology. You have the choice between pay-per-use LightOn Cloud or on-premise LightOn Appliance. Depending on the specificities of your organization and your Machine Learning usage, one might bring you more benefits for your money—our sales team is here to help you choose.

Or on-the-fly change-point detection, its use in the NEWMA algorithm allows to be 2 or 3 orders of magnitude faster than state-of-the-art approaches on typical data, and the difference would be even more significant in higher dimensions.

For Laura, the OPU is a game-changer, as fast prototyping of machine learning models for high dimensional data is becoming a reality. Now, training these models or evaluating new behaviors can run in seconds instead of minutes or even hours. During one typical working day, she might run ten small tests. Each of these computations, which used to take 20 minutes, now only takes two minutes—a three hours saving!

User testimonials

Optical Processing is one of the most innovative technologies in machine learning today. It turns what was until now mostly a beautiful theory—very high-dimensional non-linear random projections—into powerful concrete applications with tremendous computational advantages compared to classic hardware.

Maurizio Filippone, AXA Chair of Computational Statistics and Associate Professor at Eurecom, France.

For the models are simpler, and do not require much hyperparameter tuning, she is incurring little technical debt. She also knows that the models will be able to handle real-life workloads once they are in production.
Experimental particle physics at facilities like CERN generates massive amounts of data, that requires both sophisticated Machine Learning algorithms, together with some of the most advanced computing hardware.

I believe that LightOn’s OPU technology is opening a new paradigm for such extremely challenging Machine Learning pipelines, with both high data throughput and high dimensional data. My research team has selected two use cases that are currently under investigation: ‘End to End learning’ for classification of high energy proton collisions at the Large Hadron Collider; and a ‘Tracking’ task, where collisions at the LHC yield billions of records with typically 100,000 3D points corresponding to the trajectory of 10,000 particles.

Researchers in my team found the LightOn Cloud platform particularly easy to use, and within only a few weeks we have been able to obtain the first promising results. These preliminary results have just been accepted at the reference conference in particle physics ICHEP 2020.

David Rousseau, Senior Scientist at CNRS, Physicist on Higgs boson, software and AI developments.

Thousands of papers were written about the amazing potential of random projections in Machine Learning but scaling them to modern high-bandwidth data-streams (hi-res images, videos) was orders of magnitude away before LightOn’s OPU technology came along.

In my research, the usefulness of LightOn’s technology extends much further: the OPU is a perfect desktop model to study imaging through scattering media, phase retrieval, system identification, and a spate of other core imaging and signal processing questions.

Working on these topics traditionally required you to study toy numerical experiments or to have access to complicated, expensive laboratory equipment.

Being able to do _real_ experiments by simply ‘importing’ the OPU in Python and tapping this sophisticated technology from a Jupyter notebook is an incredible boon to my group’s workflow. It has been central to our research and publication process, with OPU-fueled results making it to flagship AI conferences like NeurIPS.

Ivan Dokmanić, Adjunct Associate Professor of ECE at the University of Illinois at Urbana-Champaign - Associate Professor, Department of Mathematics and Computer Science, University of Basel, Switzerland.
Optical computing: the low-hanging fruit for keeping Moore's law alive in an AI world

LightOn's ambition is to develop a hybrid electro-photonic technology for today's and tomorrow's most demanding challenges in Machine Learning. Our initial step in that foray was to build an OPU with a photonic core that could respond to some of the unmet needs in Machine Learning.

Parallel to this pathfinder effort, we have been developing an infrastructure to build a sustainable user community around that technology. The new version of LightOn Cloud opened in April 2020 featuring the latest generation Aurora 1.5 OPUs. It is available as pay-per-use, or under the LightOn Cloud for Research program—check our website for terms and conditions.

For users who prefer on-premise devices, the LightOn Appliance, including Aurora2 OPUs, has been released and is available to pre-order. You can learn more at https://lighton.ai/lighton-appliance.

Figure 16: Since its inception, LightOn's growth strategy has been to offer, through a cloud platform, early access to increasingly sophisticated prototypes. LightOn Cloud 2.0 opened on April 7th 2020 and features Aurora 1.5 OPUs. LightOn Appliance features Aurora2 OPUs, with increased performance and integration. Learn more at https://lighton.ai/lighton-appliance

A rapid development cycle

In the past four years, we have witnessed AI developments going at a much faster pace than usual semiconductor industry cycles. LightOn breaks these time constants of the chip industry, as building new AI relevant blocks go from years to months. Starting from components that are mature in the photonics industry and with known supply chains, we are building a computing pipeline that will enrich itself organically. We also expect new developments in the underlying technologies of these components to provide an additional boost at marginal cost.

In the next 3 years, our roadmap (Fig. 11) includes the development of a series of new OPUs and photonic cores, including an attendant Intellectual Property portfolio, that optically performs different mathematical functions, at scale, with speed, and energy consumption that are out of reach for pure silicon chips. Much like in the past four years, LightOn's technology is expected to continue following a x10 performance improvement every two years.

What about the Quantum Computing “elephant in the room”? At LightOn, we believe that quantum accelerators will soon find their way into hybrid processing pipelines, just like any of our OPUs. We are working on our first LightOn Quantum OPU, with integration in our cloud in 2021, while leveraging our current software stack and community.
Figure 17: LightOn’s product roadmap has two tracks: a fast iteration with customers through its LightOn Cloud service, while mature products get released as appliances for on-premise usage.

Hybrid Computing, here we come!

Guided by some of the most explosive growth in Machine Learning and Artificial Intelligence, such as AdTech or NLP, we are smoothly enlarging the current AI-electronics co-design space into a richer AI-photonics-electronics co-design space. Expect these new “software designed” computing architectures around a range of miniaturized OPU photonic cores to populate future AI acceleration boards.

Figure 18: The future of computing lies in hybrid electro-photonics co-design; here featuring several photonic cores on a single board.

The future of Moore’s law is hybrid: expect LightOn to be the leader of “beyond-pure-silicon” accelerator technology.
LightOn was founded by Igor Carron, Laurent Daudet, Sylvain Gigan and Florent Krzakala. Our team mixes a strong background in academic research as well as business development in a “deep tech” environment. LightOn has a diverse technical team of about 20 people at the cross-roads between photonics, hardware integration, DevOps, and Machine Learning.

LightOn was incorporated in 2016 and has so far raised about $5M in total. Contributors to the last $3.7M seed round included VC firms such as Quantonation (Paris), Anorak VC (San Francisco) and CEA Investissement (Paris). It has received several prizes from the Grand Prix de l’Innovation de la Ville de Paris to the AI Challenge of the Paris Region.

Web site: lighton.ai
LightOn Cloud: cloud.lighton.ai
LinkedIn: linkedin.com/company/lighton
Twitter: twitter.com/LightOnIO
Github: github.com/lightonai
Register for our monthly newsletter: mailchi.mp/lighton/nl

Additional technical insights
For additional technical insights on LightOn’s technology you may visit https://lighton.ai/our-technology