FONDAZIONE BRUNO KESSLER
MAGAZINE

The savings algorithm

September 26, 2025

FBK's Center for Cybersecurity has completed the first service of the InnovAction project, testing a GPU-as-a-Service model to optimize the cost of this valuable resource. This activity was funded through the project in collaboration with the SME Clastix. The solution works and delivers a significant impact.

As part of the InnovAction initiative—aimed at transforming Italian companies through the adoption of advanced and sustainable technologies—FBK, through its Center for Digital Industry and Center for Cybersecurity, offers services classified under the "Test Before Invest" category, as described in the official catalogue. This approach allows companies to experiment with innovative technologies in a controlled environment before fully implementing them, thereby reducing the risks associated with investment.

Thanks to InnovAction, companies can access a range of services supported by Government aid covering up to 100% of costs. The InnovAction project has received the *Seal of Excellence* under a European call from the Digital Europe Programme for the selection of European Digital Innovation Hubs (EDIHs), aimed at supporting competitiveness and growth through innovation. Running from 2023 to 2026, the project is co-financed by the Ministry of Business and Made in Italy (MIMIT), the European Union through the Next Generation EU program, and funds from the National Recovery and Resilience Plan (PNRR).

A concrete example of a service developed and implemented by FBK—funded through InnovAction—concerns the collaboration between the Center for Cybersecurity and the SME Clastix, focusing on optimizing the allocation of computational resources in the cloud computing context, with particular attention to GPUs.

GPUs (graphics processing units) are essential for the training phase of artificial intelligence techniques, which would otherwise require prohibitively long processing times using traditional CPUs. The GPU-as-a-Service (GPUaaS) model is a cloud-based solution offering on-demand access to powerful GPUs via the internet, enabling companies to leverage parallel computing power for intensive workloads such as AI, machine learning, and 3D rendering—without the need to purchase or maintain costly and complex hardware. This model allows for rapid infrastructure scaling based on demand, with a pay-as-you-go system, making high-computing technologies more accessible.

Energetic AI and the Need for Ingenuity

With the rise of Artificial Intelligence—one of the most significant technological revolutions of recent years—the growing demand for GPUs to train models and provide services to users has led to substantial increases in both operational costs and energy consumption for companies.
The energy demands of AI—particularly in the training and model creation phases—could become a major issue as these techniques are increasingly adopted. Data centers are often located near lakes (for cooling) or power plants (for energy supply). Reducing GPU-related energy consumption is therefore key to ensuring that the benefits of AI do not come at the expense of environmental sustainability.

Problem Solving

The project focused on designing an **optimization algorithm** to improve GPU usage efficiency, aiming to reduce operational costs while maintaining the required quality of service.

Specifically, the algorithm enables better optimization of GPU resources used by AI applications (such as Large Language Models), lowering the overall cost for those renting GPUs from cloud providers.

The problem resembles a bin-packing challenge: finding the best way to fit the maximum number of objects (AI applications) into the smallest number—or the least expensive—containers (GPUs). Unlike a simple bin-packing problem, however, this scenario adds the complexity of maintaining high service quality, without degraded performance.

Silvio Cretti and Marco Zambianco, from the Distributed AI for dependent cyberSecuritY (**DAISY**) Unit of the Center for Cybersecurity, worked on the project, which concluded in July 2025.

Bridging the Gap

"The greatest difficulty," commented Cretti*, "was understanding each other and being able to speak the same language. Often, the company's pragmatic approach clashes with the more theoretical orientation of the research center. The jargon differs—sometimes the same terms mean different things in different contexts, or a concept may exist in one domain but not in another (e.g., simulation, optimization, orchestration). Typically, a company is looking to solve a concrete problem (e.g., cost reduction in a specific context), whereas a research center may offer a general solution not easily adapted to that context. The challenge lies* ***in adaptation—essentially, a 'problem looking for a solution' and a 'solution looking for a problem'*** *must meet. In this project, we managed to find a compromise—not perfect for either side, but good enough for both."*

Mutual Benefits

"The research center," added Cretti*, "was able to demonstrate that its solutions work—not just in theory or in simulation—but in a context resembling a real operational environment. The company, meanwhile, received a proof of concept*

showing that the cost of using GPUs can be mitigated through optimization: ***with the algorithm developed by FBK, a theoretical savings of around 20% in operational costs can be achieved while maintaining adequate performance****."*

"Energy management algorithms for GPUs like the one developed in this project," **Silvio Ranise**, Director of the FBK Center for Cybersecurity, *said "can also enhance the security of the AI training phase. They act like a traffic controller, preventing any single process from monopolizing resources, thereby reducing the risk of denial-of-service attacks that can slow or stop computation. Another benefit is that these algorithms smooth out GPU power consumption, making it harder for attackers to analyze consumption patterns and infer training data or model behavior—helping to defend against power side-channel attacks."*

Cover foto: Silvio Cretti and Marco Zambianco, Center for Cybersecurity at FBK

PERMALINK

https://magazine.fbk.eu/en/news/the-savings-algorithm/

TAGS

- #ai
- #artificialintelligence
- #clastix
- #cloud computing
- #cybersecurity
- #daisy
- #digitalindustry
- #edih
- #GPU
- #GPUaaS
- #innovaction
- #llm
- #Machine learning
- #pnrr
- #proofofconcept
- #sme
- #testbeforeinvest