# Innovative Expansion of HPC Infrastructure for Scalable AI Inference Using MACx GPUs

Tomi Ilijaš, Tristan Pahor and Tomislav Šubić
Arctur d.o.o., Nova Gorica, Slovenia

European high-performance computing (HPC) environments are increasingly challenged by the rapid growth of artificial intelligence inference workloads. While traditional GPU-accelerated HPC systems were primarily optimized for simulation and training tasks, the emergence of large-scale language models and real-time AI services requires a new approach that balances performance, cost efficiency, and architectural openness. This work presents an innovative strategy developed by Arctur for expanding an existing HPC infrastructure with more than 100 MACx GPUs, creating a scalable inference platform that reduces dependency on proprietary ecosystems and addresses vendor lock-in challenges associated with CUDA-centric deployments.


The proposed architecture integrates MACx accelerators into an established HPC environment through a hybrid scheduling layer that enables both traditional HPC workloads and AI inference pipelines to coexist. The MACx platform provides a general-purpose GPGPU architecture supporting FP64, FP32, TF32, BF16, FP16, and INT8 precision modes, allowing seamless execution of scientific computing and AI frameworks such as PyTorch and TensorFlow. By leveraging open-source toolchains and transparent code migration strategies, Arctur achieved rapid deployment without significant refactoring of existing applications. The integration emphasizes modular cluster design, enabling incremental scaling and flexible resource allocation for enterprise and research use cases.

Initial deployment demonstrates that large inference clusters built on non-CUDA accelerators can achieve competitive throughput and predictable operational costs compared to traditional inference nodes. Benchmark observations indicate that the MACx-based infrastructure enables efficient batching of language-model inference workloads while maintaining compatibility with existing HPC data pipelines. Furthermore, the heterogeneous scheduling approach reduces idle GPU time and improves overall system utilization across mixed workloads.

The presented deployment highlights a practical pathway toward sovereign and open AI infrastructure in Europe. By combining HPC expertise with alternative accelerator technologies, Arctur demonstrates that vendor diversification can be achieved without compromising performance or developer productivity. The approach is particularly relevant for European organizations seeking to expand AI capabilities while maintaining strategic autonomy and long-term cost sustainability. Future work will focus on

optimization of inference runtimes, deeper integration with EuroHPC workflows, and evaluation of large-scale multi-node inference scenarios.

References

[1] European High Performance Computing Joint Undertaking, AI Factories Initiative, European Commission (2024–2026).

[2] Garcia-Lopez, P., Barcelona Pons, D., Copik, M., et al., AI Factories: It's Time to Rethink the Cloud-HPC Divide, arXiv (2025).

[3] Turisini, M., Amati, G., Cestari, M., LEONARDO: A Pan-European Pre-Exascale Supercomputer for HPC and AI Applications, arXiv (2023).

[4] Nikolic,S., Filipovic L., Ilijas T., Vukotic M.: FIT4HPC? -Accelerating digital transformation by supercomputing opportunities, The Journal of Supercomputing (2025)