Case Study

intel

Intel and VNClagoon -
High-speed GenAI processing with security by design

The challenge: Most generative AI models nowadays run in large cloud environments, which makes them vulnerable to malicious attacks. It's time to run them on your local or closed systems, where you can safely integrate your data and get the best performance out of your own algorithms – on a deep-state encryption and hardware level, utilizing your own PCs and your own data centers.

Learn more about this collaboration between Intel and VNClagoon.

GenAI is changing the way the world works

– it makes us even more relying on data as fuel for success in the business world.

Generative AI is promising exponential growth in solving many data challenges.

More and more companies are piloting programs and integrating GenAI into their business workflows, setting high demands on the underlying processing capabilities:

- Public sector companies want to streamline administrative processes with millions of different data points
- Banks analyze thousands of different transactions in real-time to prevent fraud
- Consumer goods companies want to target millions of customers with individual messages

In a world run on data, tasks that seemed impossible to solve a few years ago are suddenly manageable. But there are also risks at stake when using GenAI – it has a high data demand – to get the optimal outputs, companies need to share sensitive company, client, and customer data.

Especially for companies that want to keep their customer and internal data secure, but still want to benefit from the newest advances in generative AI technology, using all the software packages that a modern workplace needs: It's time to utilize the technology in a secure and safe sandbox environment. Welcome, VNClagoon!

Who is VNC?

VNC (VNC - Virtual Network Consult AG) is a global software corporation that develops open source-based communication and collaboration applications, with headquarters in Zug, Switzerland. VNClagoon is their specialized secure communication and collaboration suite.

Their greatest strength is that VNClagoon is entirely based on open-source technology and allows any operations to be on site or within a secure cloud environment.

VNClagoon follows a modular architecture and offers a secure unified communication & collaboration suite including:

- Groupware (email, calendar, contacts)
- Chat and videoconferencing
- CRM
- Task, project management
- File management
- Directory
- Dashboards

All these features come with an integrated **modern generative AI assistant that enhances communication and productivity, powered by Intel's latest processing technology.**

This makes VNClagoon an ideal solution for large enterprises – in healthcare, public sectors, financial services – that handle sensitive data.

Learn more here: www.vnclagoon.com

Confidential AI with the Intel technology stack –

We found the perfect match for your GenAI challenges!

Taking the best of both worlds – Intel's latest high-speed processors – and the safe VNClagoon environment, we provide you with top-notch performance and increased security layers. The solution utilizes the Intel OpenVINO™ toolkit that significantly reduces inferencing latency, thus improving execution time, while maintaining high-quality outputs – in the cloud or on your local device.

Combined with VNClagoon's privacy suite, Intel technologies allow you to safely run GenAI algorithms, particularly large language models (LLMs) within a safe sandbox environment.

Through our business collaboration, GenAI applications are decoupled from the platforms of the large AI solution providers – preventing any unauthorized access. This system works best on PCs powered by the latest generation of Intel® Core™ Ultra processors where you will benefit from improved generative AI performance and power efficiency.

This technological combination comes with some additional key benefits:

Staying in charge of your training data – anywhere and anytime

Using the Intel OpenVINO toolkit, you can deploy AI algorithms within the VNClagoon environment. Any training data for GenAI can be securely run, continuously improving it with every iteration. Many LLM code bases rely on a similar framework, but by running them with Intel OpenVINO you can highly accelerate their run-time. The toolkit improves the inference latency of an AI model, enabling high-speed operations in secure cloud environments.

Making the most of your privacy

Are you looking for a safe framework to access and securely process data? VNClagoon is a safe sandbox environment where you can store, merge, and update your data. The suite significantly minimizes the threat of unauthorized access or data breaches. Powered by an Intel® Xeon®-based-cloud infrastructure featuring Intel SGX and being accessed from a PC based on Intel Meteor Lake architecture, we can guarantee a maximum level of privacy and security.

Truly master AI compliance

Regulators set high demands on data security, and future regulations (such as the European Union's AI Act) might hinder advances in GenAI tech within your software stack. Utilizing GenAI in a fast-paced world constantly leads to new industry standards. The collaboration between VNClagoon and Intel puts you ahead when dealing with sensitive customer and client data – making your software stack future-proof.

Technical details for Intel OpenVINO™

(Open Visual Inference and Neural Network Optimization)

The Intel OpenVINO™ technology is optimized for low-resource environments. It runs on end-to-end secure devices using laptops with the newest Intel® Core™ Ultra Meteor Lake chips but also reduces latency whenever a GenAI agent is accessed from a cloud environment

- OpenVINO allows you to map pre-trained generative AI algorithms to smaller models with a fixed set of training data. It uses a model optimizer (command-line tool) that helps you transition between a training and a deployment environment
- OpenVINO runs flexibly on different hardware setups (on GPUs, CPUs, and NPs; on-premise and on-device; in the browser, or in the cloud)
- Using OpenVINO's static model analysis for optimal execution helps you to significantly boost speed for any down-scaled generative AI applications
- OpenVINO optimizes models and converts them to intermediate representation (including XML descriptions, .bin file with "weights", and binary data). In later stages, it passes them on to an inference engine to further optimize the algorithms based on the chosen target end device
- This process, on average, reduces the size of a **GenAI model by about 50 percent**! The great thing: This also reduces runtime memory during compiling, making it possible to run algorithms on edge systems instead of large data centers

Secure Cloud – Protecting data models with Intel confidential computing

Intel® Xeon® is a scalable processor architecture ideally suited for large data centers, offering cloud architecture that runs on LLMs such as the ones utilized by VNClagoon.

The Intel® Xeon® CPUs come readily equipped with Confidential Computing through both Intel SGX (Software Guard Extensions) and Intel TDX (Trust Domain Extensions). Both technologies create Trusted Execution Environments (TEE) that offer a couple of key advantages for generative AI technology:

- Intel Confidential Computing is working directly on the hardware level. Data processes that run with Intel SGX can use the enclave-like process separation, making it impossible for them to be read out by external code (even higher-privileged code, such as OS code and other hardware infrastructure), while Intel TDX provides full virtual machines inside the TEE, separating workload owner from infrastructure operator.
- Both Confidential Computing technologies enable consistent auditing and hardware security update checks on your servers. Before any data sharing between different system parts, there is a specified attestation mechanism (cryptographically verifying that the TEE is genuine and configured as expected), so you know that your applications have not been compromised
- Through these security layers, your data in the server backend as well as your data flow to your client-based generative AI models are always secure!

Conclusion: The next level of generative AI computing is waiting for you!

VNClagoon (combined with Intel OpenVINO™ and Intel Confidential Computing) is the perfect fit to elevate your communication and collaboration efforts. Apart from the generative AI security layers, the combination offers a couple of key capabilities:

- Real-time content indexing from any VNClagoon product powered by Apache Solr enabling smart search and chatbots
- Multi-device frontend (including edge computing on end-user devices)
- Current integration of Meta's Llama in VNClagoon (with eight more AI models coming soon). The collaboration offers the LLMs to run safely on your local systems, protecting your data used to train the LLMs as well as the output generated by them and the proprietary models themselves

Together, Intel and VNClagoon help you get the most out of your hardware infrastructure! Just contact our experts or reach out to your preferred Intel resellers or integrators to set up your own VNClagoon with Intel® Core™ Ultra Meteor Lake processors, utilizing Intel OpenVINO™ and building on top of backend services powered by Intel® Xeon™ processors and protected by Intel SGX today and Intel TDX tomorrow.

Discover more, and watch the full demo here: New VNClagoon AI Demo Video - VNClagoon

Customer reference

Based on open-source technology and utilizing Intel Confidential Computing, the VNClagoon solution is currently being developed and rolled out as the secure AI frontend for the eArchive4Future project led by Fujitsu. This solution aims to provide a sovereign technological path into the future for Germany's largest public sector document archive, employed by the German Federal Employment Agency (Bundesagentur für Arbeit).

Intel is committed to respecting human rights and avoiding complicity in human rights abuses. See Intel's Global Human Rights Principles. Intel's products and software are intended only to be used in applications that do not cause or contribute to a violation of an internationally recognized human right.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.