Solution Brief

Small and Medium Sized Enterprises and Branch Offices

CASwell and Intel Deliver AI Server for Private LLMs

Caswell CAR-5071 has dual 5th Gen Intel® Xeon® Scalable processors and optional Intel® Arc™ GPUs for the Edge offering the performance needed for AI large language model-based inferencing applications

The market for private AI large language model (LLM) solutions is expanding as enterprises across industries recognize the value of leveraging advanced AI capabilities while maintaining control over their data. A private AI solution enables organizations to host and manage AI models on-premises or within a secured cloud environment, ensuring data sovereignty and reducing the risks associated with third-party data exposure.

The demand for these services stems from growing concerns about data privacy, compliance with governmental regulations like GDPR and HIPAA, and the desire to take advantage of the capabilities of LLMs.

The use cases for private LLMs span a wide range of industries and applications. In healthcare, for example, organizations can use private AI to analyze patient data, generate clinical insights, and support decision-making without violating patient confidentiality.

In the finance industry, models can assist with fraud detection, customer support, and investment analysis while complying with stringent regulatory requirements. Retail and e-commerce companies can leverage private AI for personalized customer experiences, inventory optimization, and demand forecasting. Additionally, manufacturing firms can deploy AI for predictive maintenance, supply chain optimization, and quality control.

Across all these sectors, private AI provides the flexibility to tailor models to specific workflows and integrate them seamlessly into existing systems.

To support efficient inferencing, a private AI server must meet several key technical requirements. First, it should offer high-performance Intel® architecture computing capabilities, optionally with GPUs optimized for AI workloads, to handle the complex operations involved in processing large-scale language models.

Second, the system must include ample data storage capacity to accommodate the significant data volumes required for model fine-tuning and inferencing. Low-latency networking is another critical component, enabling real-time responses for applications like conversational AI and recommendation systems.

Moreover, the server should incorporate advanced network and security features, such as encryption, role-based access control, and audit logging, to protect sensitive data and ensure compliance with industry standards.

Finally, usability and scalability are essential considerations for private AI servers. Organizations need user-friendly interfaces and APIs that allow seamless integration with existing infrastructure and workflows. The system should also support scalability, enabling businesses to expand their AI capabilities as their needs grow.

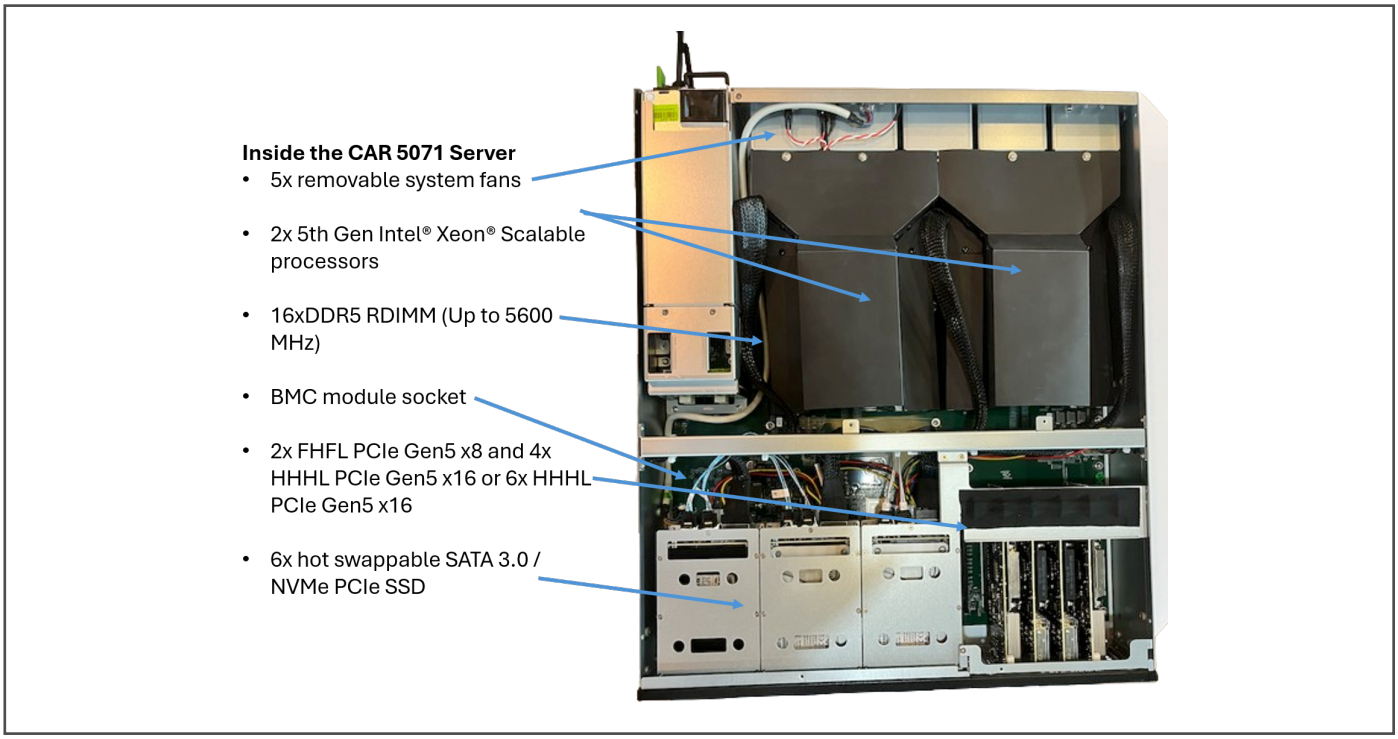

Figure 1. Block diagram of CAR-5071 server.

With its CAR-5071 server, CASwell, an Intel® Industry Solution Builders' Network Builders Community member, has built a server with the features and capabilities needed for private AI applications.

CAR-5071 is Designed for Private AI at the Edge

CASwell's edge private AI system is the CAR-5071 (see Figure 1), a high-performance multi-networking rackmount server that features dual 5th Gen Intel® Xeon® Scalable processors and can support up to two Intel® Arc™ GPUs for the Edge for focused AI processing. This provides the compute power to run both the AI inferencing workloads with left over cores that can be used for cyber security services.

The system features six PCIe x16 standard slots with support for half-high, half-length cards that interchangeably support network and storage modules. Additional dedicated storage support is available via six swappable U.2/SATA SSD (PCIex4) slots.

The CAR-5071 incorporates six internal PCIe Gen 5 slots that can also be equipped with GPUs or accelerators such as advanced SmartNICs. For ECC main memory, the CASwell CAR-5071 accommodates up to 16 DIMMs of DDR5-5600 memory for a total of up to 512GB.

5th Gen Intel® Xeon® Scalable Processors Provide AI Performance

5th Gen Intel Xeon Scalable processors deliver increased performance per watt and lower total cost of ownership (TCO) across critical edge and data center workloads including artificial intelligence and security.

These processors feature AI acceleration in every core in order to address demanding end-to-end AI workloads without the need for discrete accelerators. These CPUs feature up to 42% higher inference performance and less than 100 millisecond latency on large language models (LLMs) under 20 billion parameters[1].

Building on previous generations, 5th Gen Intel Xeon Scalable processors bring new innovations to deliver performance and efficiency benefits to customers. The processors support up to 64 cores per socket or processor and nearly three times the maximum last-level cache from the previous generation. They offer eight channels of DDR5 per CPU, support DDR5 at up to 5,600 megatransfers per second (MT/s), and increase inter-socket bandwidth with Intel UPI 2.0, offering up to 20 gigatransfers per second (GT/s). 5th Gen Intel Xeon Scalable processors are pin-compatible with the previous-generation 4th Gen Intel® Xeon® Scalable processors.

Intel® Arc™ GPU for the Edge

The other Intel component that CASwell standardized on was the Intel Arc GPU for the Edge. Intel Arc GPU for the Edge is based on the highly scalable Intel® X^{e}-core architecture and designed to enable innovation for AI, visual computing, and media processing.

Intel X^{e}-core architecture represents a significant evolution in GPU design, emphasizing versatility and performance across a spectrum of computing needs from cloud to edge. It incorporates advanced features such as scalable data parallelism, efficient AI acceleration and support for rich graphical rendering techniques. This architecture enables Intel Arc GPUs for the Edge to deliver high performance for AI applications.

Intel Arc GPUs for the Edge target the edge specifically with five-year long-life availability and support, diverse edge-focused form factors and support for edge-constrained usage conditions.

With support for a more open, standards-based software stack, customers can run high-performance AI applications and solutions using the Intel® Distribution of the OpenVINO™ toolkit. OpenVINO software tools provide a streamlined development workflow to deploy inference workloads.

Private AI Solution Featuring Ollama

To demonstrate the performance and features of the server, CASwell built a Docker-based cloud native private AI edge server running Windows Server 2022 using the Intel® Xeon® Gold 6548Y+ processor with 32 cores.

The CASwell private AI solution was configured to run Ollama, an open-source tool that enables developers to develop highly responsive AI-driven chatbots running entirely on local servers.

Ollama supports many LLM models that are optimized for applications like natural language processing (NLP), code generation, data analysis, text and image processing, scientific applications and more. With no reliance on external servers, Ollama reduces processing latency making the models faster and more reliable.

The CASwell demonstration solution is designed for the following applications:

Cybersecurity

- Helps translate cyberthreat data and analysis into natural language, enabling analysts with fewer technical skills to work more efficiently.
- Uses generative AI to identify remediation steps, allowing new team members to quickly learn and effectively respond to cyberattacks.

Smart Factory

- Implements an intelligent edge solution for sensor-based predictive maintenance, detecting potential failures and maximizing system uptime.
- Supports quality control applications to monitor product quality and optimize processes without exposing sensitive manufacturing data.

Conclusion

Private AI has found a growing market of companies that need the capabilities and benefits of AI LLM services delivered from an onsite server that can reduce the risk of exposing valuable or sensitive data. In its CAR-5071, CASwell has provided for all of the needs of a private AI server with dual 5th Gen Intel Xeon Scalable processors and up to two Intel Arc GPUs for the Edge.

The company has demonstrated the ability of these systems configured with the open source Ollama AI chatbot development tool. The server was able to meet the needs of key use cases in education and manufacturing.

CASwell's CAR-5071 shows how server design using Intel AI components can develop a complete server solution offering the features and performance needed to help companies benefit from private AI.

Learn More

CASwell homepage

CASwell CAR-5071

Unlock the Power of Edge AI Computing with CASwell

Intel® Xeon® Scalable Processors

Intel® Arc™ GPU for the Edge

Intel® Industry Solution Builders

Intel® Distribution of the OpenVINO™ toolkit

[1] https://www.intel.com/content/www/us/en/newsroom/news/5th-gen-xeon-data-center-news.html#gs.jtna5l

Notices & Disclaimers

Performance varies by use, configuration and other factors.

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

Intel optimizations, for Intel compilers or other products, may not optimize to the same degree for non-Intel products.

Your costs and results may vary.

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

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

See our complete legal Notices and Disclaimers.

Intel is committed to respecting human rights and avoiding causing or contributing to adverse impacts on human rights. 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 adverse impacts on human rights.

0325/AL/DJA/PDF Please Recycle 364770-001US