Technology | United States | SME | Agentic AI

Slashing the cost of AI-powered local government services

Onetera teams up with Lenovo to make its local government AI solutions available on-premises, dramatically cutting the cost of running token-intensive workloads.

Headquartered in Los Angeles, Onetera provides agentic AI solutions for local governments of all sizes—from small towns to big cities to entire counties. To support the different needs of these different jurisdictions, Onetera offers the flexibility to deploy its AI-native government platform across both cloud and on-premises environments.

Eager to enable its municipal customers to benefit from AI in their own protected environments and to drive down AI workload processing costs, Onetera worked with the Lenovo AI Center of Excellence (CoE) to validate the accelerated breakeven and favorable token economics of owned on-premises infrastructure.

Running Onetera on-premises gives local governments:

- Cost-efficient deployment path for agentic AI
- Peace of mind of running workloads on their own hardware
- Complete data control
- Reliable performance

98%

cost saving for document vision analysis

90%

cost saving for document OCR

65%

cost saving for FAQ response generation

- Full-stack solution offering in partnership with Lenovo
- Huge cost advantage over cloud APIs
- Full data sovereignty

NVIDIA

ONETERA

The challenge

Smarter municipal operations on a budget

Faced with shrinking budgets and growing demand for services, local governments are under constant pressure to do more with less. Onetera leverages agentic AI to strengthen connections between local governments and their communities, streamlining access to resources.

Managing tokenomics and AI expenses

- Government agentic AI use cases typically involve multi-step reasoning, retrieval, and structured output generation, driving token consumption well above simple chat. In fact, a full document processing run can generate upwards of 72,000 tokens.
- Large token volumes and high cloud API pricing prompted Onetera to investigate on-premises compute as a deployment option for its customers.
- Data sovereignty is another key concern. Many state and local data governance policies prohibit organizations from moving sensitive personal information to third-party cloud inference endpoints.

The solution

Making on-premises infrastructure an option

Onetera engaged with the Lenovo AI CoE to test out its production inference layer with on-premises hardware for the first time.

AI-optimized hardware and technical expertise

- Data scientists from the Lenovo AI CoE helped Onetera to deploy the large language models (LLMs) on a Lenovo ThinkSystem SR675 V3 server, swapping out the tool calls from third-party cloud providers to the self-hosted LLMs.
- Every stage of the pipeline, from document ingestion to generation, was kept inside the infrastructure boundary.
- Working closely with the Lenovo AI CoE team, Onetera validated multiple workload types during a 40-day pilot and completed a thorough stress test.
- During the pilot period, only one of the server's eight available NVIDIA RTX PRO 6000 GPUs was actively used, leaving 672 GB of VRAM as headroom for scaling.

Lenovo hardware and software

Hardware:

- Lenovo ThinkSystem SR675 V3 Server
- NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs

Software:

- NVIDIA AI Enterprise
- NVIDIA CUDA Toolkit
- NVIDIA Nemotron
- Ubuntu

> **"To deploy really large models, you need really powerful compute, and the GPU-accelerated Lenovo ThinkSystem SR675 V3 delivers just that. We ended up using only around 30% of the entire system for one of our largest models, and were able to comfortably run our analyses."**

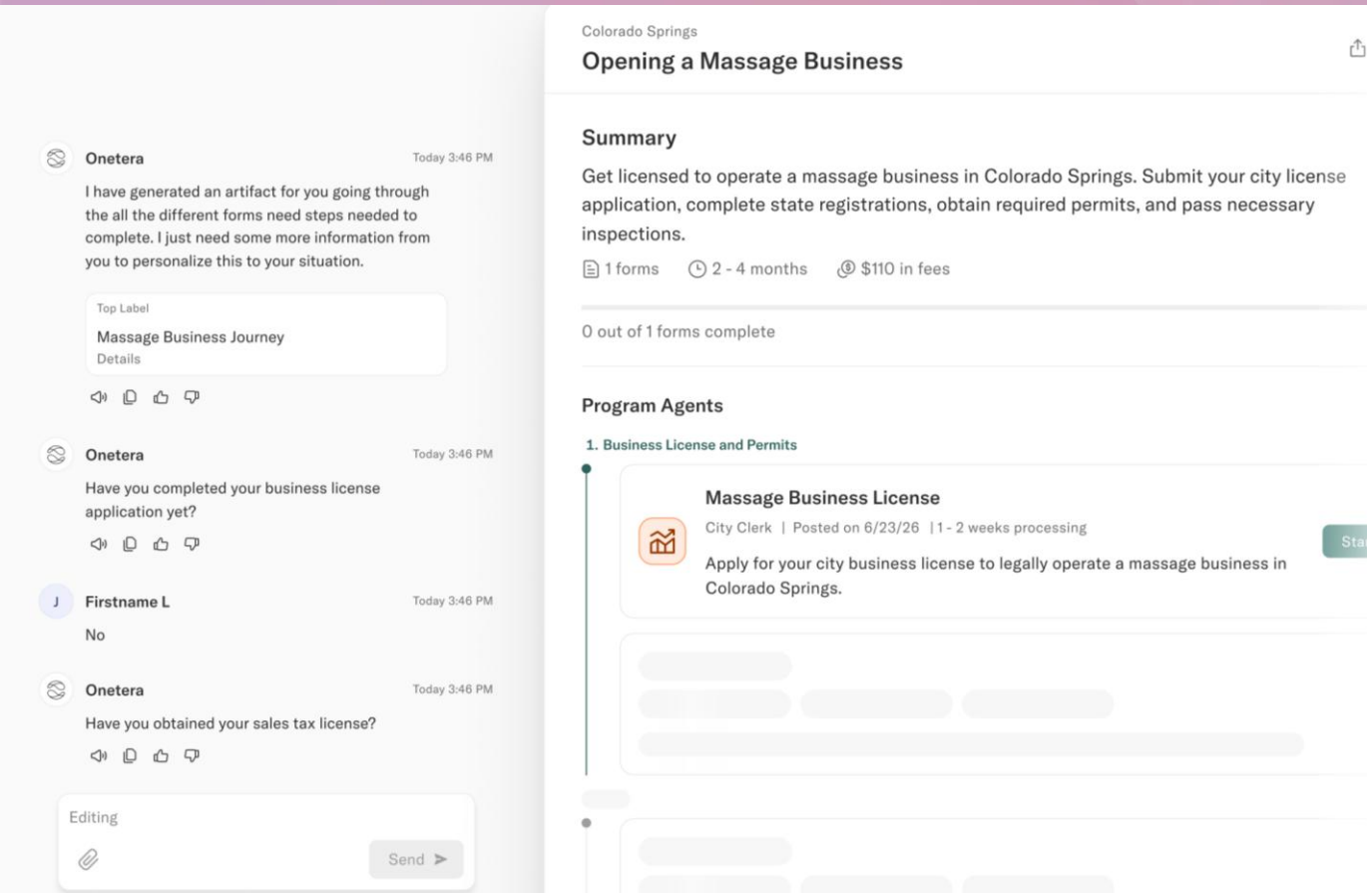
Sumukh Aradhya

Software Engineer, Onetera

30 to 80
concurrent requests with 20 to 50 users

Up to 111k
tokens input and up to 16k tokens output

<10s
hard latency and up to 50 tokens per second throughput



The outcome

Cutting costs while improving security

The results of the validation pilot have been phenomenal, with Onetera achieving cost savings of up to 98% for token-heavy workloads such as document vision analysis running on the Lenovo ThinkSystem SR675 V3 server.

“This was only during the course of a short 40-day trial; once you scale up, the cost savings are only going to increase further,” comments Sumukh Aradhya, Software Engineer at Onetera. “Newer, more expensive cloud LLMs are being released all the time. So, doing the same thing on premises with locally hosted models is only going to become more economical.”

As well as cutting costs, deploying Onetera’s AI-native government platform on-premises ensures complete control over where data is stored and where data flows. “It is a workload where on-premises infrastructure is not merely the cheaper option but the only compliant one,” confirms Felix Ruano, Founder & CEO of Onetera.

The impact is clear:

- **Up to 98% cost savings** makes agentic AI more accessible to local governments, helping to accelerate their AI journey
- **Full data sovereignty and control** ensure compliance with state and local data governance policies
- On-premises infrastructure delivers **reliable performance and the peace of mind of physical control**

By going to market with Lenovo, Onetera gives local governments the option of deploying agentic AI on premises, or in a multi-cloud approach—unlocking new opportunities across for local government customers to have their AI compute and data where they want it.

Lenovo AI Center of Excellence

Hands-on, expert guidance for AI success

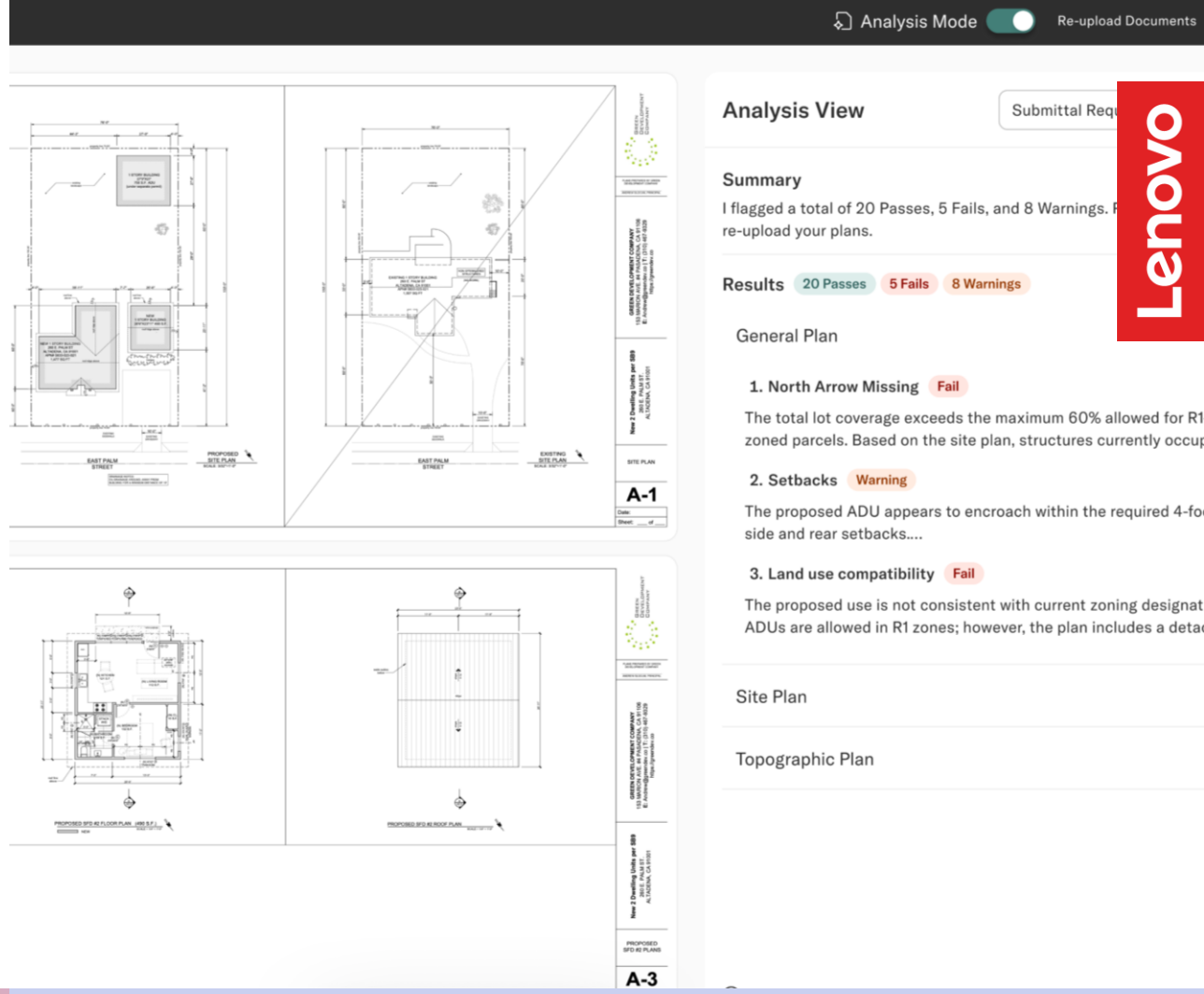
Onetera's success in cutting the cost of running agentic AI workloads by moving from the cloud to on-premises infrastructure demonstrates how the Lenovo AI CoE can help your organization gain a competitive advantage.

What's available:

- **AI-optimized infrastructure** for testing and validating software
- **Data scientists and technical engineers** with the expertise and insights to support your goals
- **AI Innovators partner ecosystem** provides tailored solutions for end-to-end operations, including computer vision, audio recognition, prediction, security, and virtual assistants
- **Over 150 AI-ready solutions** from the AI-Ready Solutions Program

Lenovo's AI Services specialize in helping businesses of all sizes navigate the AI landscape. From AI Discovery to AI Fast Start, the Lenovo AI CoE experts will meet you where you are and take AI from concept to reality.

Learn more



60-70%

cost saving through owning a Lenovo ThinkSystem SR675 V3 server versus renting equivalent dedicated GPU capacity in the cloud

Want to know more?

Get in touch with Eric Skomra (eskomra@lenovo.com) to set up a chat.

Lenovo 2026 Lenovo Internal. All rights reserved.