Potential Blog: Letta Thoughts

Running Letta for a Cloud application  
Letta Cloud offers several pricing tiers, including a generous free plan, with the cost depending on your usage of model requests, active agents, and storage. For higher volumes and advanced features, there are Pro and Scale plans.

Subscription plans

* **Free Plan:** This tier is for users getting started with Letta Cloud.
  + **Cost:** $0 per month.
  + **Limits:** 50 premium requests, 500 standard requests, 100 active agents, 2 agent templates, and 1 GB of storage.
* **Pro Plan:** Designed for shipping agents in production.
  + **Cost:** $20 per month.
  + **Limits:** 500 premium requests, 5,000 standard requests, 10,000 active agents, 20 agent templates, and 10 GB of storage.
* **Scale Plan:** Intended for teams deploying agents at scale.
  + **Cost:** $750 per month.
  + **Limits:** 5,000 premium requests, 50,000 standard requests, 10 million active agents, 100 agent templates, and 100 GB of storage.
* **Enterprise Plan:** For organizations with high-volume or specialized needs.
  + **Cost:** Custom pricing based on requirements.
  + **Features:** Can include unlimited agents and storage, custom model deployments, and dedicated support.

Standard vs. Premium requests

Letta Cloud distinguishes between two types of model requests. These are priced differently and have separate monthly limits within each plan.

* **Standard models:** These are faster and more economical, suitable for basic chat and simple tool-calling tasks. Examples include GPT-4o mini and Gemini Flash.
* **Premium models:** These offer enhanced capabilities for complex tasks and are used for advanced reasoning. Examples include GPT-4.1 and Claude Sonnet.

Additional and usage-based costs

* **Usage-based pricing:** Users on the Pro plan can continue making model requests after exceeding their monthly quota by enabling usage-based pricing.
* **Potential hidden costs:** Some analyses suggest that enterprise and usage-based pricing on Letta may lack transparency. This can complicate budgeting for scaling organizations.
* **AWS infrastructure costs:** Additional infrastructure costs from Amazon Web Services may be incurred on top of the Letta subscription when running Letta on an AWS Marketplace product.

# Running Letta on AWS Fargate

Yes, you can run the Letta server on AWS Fargate. Letta is designed to be deployed in containers, and AWS Fargate is a serverless compute engine for running containers with Amazon Elastic Container Service (ECS).

General deployment on Fargate

The process involves using the official Letta Docker image (letta/letta:latest) and deploying it as an Amazon ECS task.

Steps for deployment

1. **Create an ECS cluster**: Set up a new ECS cluster to host your Letta server.
2. **Define a Task Definition**:
   * Specify the letta/letta:latest image.
   * Map the container port 8283 to a host port to expose the Letta server.
   * Set environment variables in your task definition for configuring the Letta server, such as OPENAI\_API\_KEY, SECURE=true, and LETTA\_SERVER\_PASSWORD, to integrate with an LLM and protect your instance.
3. **Create a Fargate service**: Use the task definition to create a Fargate service that will run and maintain the desired number of Letta tasks.
4. **Add a load balancer**: For production, create an Application Load Balancer (ALB) to route traffic to your Fargate service. This allows for scalability and high availability.
5. **Use a database**: For persistent storage of your agent data, Letta can use a PostgreSQL database. Set the appropriate environment variables and storage volumes in your task definition to connect your Letta container to an AWS managed PostgreSQL instance (like Amazon RDS).