**Air Cloud Documentation (Updated English Version) /** 25.05.27 17:50 Updated

**1. Overview**

**Air Cloud** is a cloud computing platform built to run AI inference at scale. Whether you're deploying for real-time AI services or cost-sensitive workloads, Air Cloud gives you two options:

* **Air Cloud**: A distributed, crowd-sourced GPU infrastructure offering the lowest possible pricing through our trusted GPU network.
* **Air Cloud+**: Proprietary GPU devices operated by AIEEV, offering guaranteed uptime, performance, and availability.

Forget complex configurations and overpriced legacy cloud platforms. Traditional bare-metal GPU clouds often fall short in dynamic, production-grade environments — Air Cloud is purpose-built to solve that.

[Get started instantly by creating an account.](https://integrated-dev.aieev.cloud:3002/ko)

**A. Choose Your Deployment Method**

* **Air Container**: Seamlessly deploy containers from public or private repositories (AWS, GitHub, Docker Hub, Google). Just enter your image address and secure token, and you're good to go.
* **Air Model** *(Coming Q3 2025)*: Designed for faster AI inference with serverless, pay-per-second billing. Automatically pulls your models, sets up REST endpoints, and enables autoscaling for your production workloads.
* **Air API** *(Coming Q3 2025)*: Offers OpenAI-compatible REST API endpoints using open-source models like Whisper, LLaMA 3.3 8B, Solar Mini, and D.N.A. Features serverless.

**B. Choose Your Infrastructure**

* **Air Cloud**: A distributed network of over 10,000 GPUs in Korea, sourced from our secure peer-to-peer partner network. We’ve pre-secured 1,000+ GPUs that can be dynamically assigned to your job at ultra-competitive pricing.
* **Air Cloud+**: AIEEV-owned GPU servers designed specifically for AI inference, offering high availability, service stability, and enterprise-grade security.

**C. Our Mission**

We make cloud computing accessible and affordable for everyone — without compromising performance, reliability, or usability. Our goal is to empower both individuals and enterprises to unlock the full potential of AI and scalable infrastructure.

**Need help?** Check our documentation, reach out via Discord([Discord](https://discord.com/channels/912829806415085598/1187492973148115076/threads/1349672733168107541)), chat support, or email(contact@aieev.com). Details are on our Contact page.

**D. What’s Next?**

* Run Air Cloud Tutorials
* Add funds to your account

**2. Get Started**

Welcome to Air Cloud. This step-by-step guide will get you up and running with your first deployment in under 5 minutes.

To launch your resources, you’ll need both a billing account and at least one project. The onboarding flow in the console guides you through this.

**A. Creating an Account**

**텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.**

**Figure 1. Sign Up and Log In**

Start by signing up for an Air Cloud account to manage and access your deployments. You can sign up using GitHub, Google, or Email ID. Once registered, log in to begin using the platform.

**B. Adding Funds**

New users receive $20 in credit upon registration. To deploy resources, you must fund your account. Currently, bank transfers and credit card payments are supported.

To load your balance, email us at [sales@aieev.com](mailto:sales@aieev.com) — our team will respond immediately. Bulk credits and reserved-resource discounts are available.

**C. Creating a Project**

**텍스트, 스크린샷이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.**

Figure 2. Create Project

Once your account is funded, create a project to deploy and organize your workloads. You can deploy multiple containers, models, or APIs within a single project to match your AI service architecture.

**D. Managing Your Team**

You can either create a personal account or get invited by a team member.

멀티미디어 소프트웨어, 스크린샷이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

Figure 3. Organization Members Page

**텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.**

**Figure 4. Invite Member**

* **Sign Up**: Use GitHub, Google, or Email ID.
* **Invite Team Members**: Admin users can invite teammates via email through the Member Management page. Team members must register using the exact email address they were invited with. After registration, they must accept the invite to join the organization and project.
* **Role Types**: Admins can manage team settings, invite/delete members, and remove projects. Users have limited access to deployment and monitoring tools only.

Note: If a user receives an invite, ensure they log out from all accounts before accepting the invitation to avoid errors (e.g., 404 page).

**E. Billing Information**

* **Air Container** and **Air Model** are billed hourly. For full pricing details, see our **Billing FAQ.**
* **Payment Method**: supports local and international wire transfers as well as credit card payments.

**3. Air Container**

Air Container enables you to deploy containerized AI services quickly and efficiently. Here's how it works:

**[Base Settings]**

**멀티미디어 소프트웨어, 스크린샷이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.**

**Figure 5. Enter General Information**

스크린샷, 텍스트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

**Figure 6. Enter Image Information**

During container creation, the following fields are required:

* **Name**: A user-friendly name for the container.
* **Category**: Type of service the container provides (used for future Playground integration).
* **Container Image**: Docker image URL of the container.
* **Registry Provider**: The registry where the image is hosted (e.g., GitHub Container Registry, Docker Hub).
* **Registry Username & Password (optional)**: Required if pulling from a private registry.

**[Advanced Settings]**

**스크린샷, 텍스트, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.**

**Figure 7. Advanced Settings**

In the "Advanced" section, additional options are available:

* **Start Command**: Overrides the default startup command in the container image.
* **Port**: Overrides the default exposed port.
* **Health Check URL**: Path to check container health status (e.g., /api/health).
* **Environment Variables**: Set key-value pairs required for your app (e.g., DB credentials, API keys).

**[Deployment Completion]**

**텍스트, 소프트웨어, 멀티미디어 소프트웨어, 스크린샷이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.**

**Figure 8. Deployed Container Information**

Once deployed, the container will start immediately, and an API endpoint or service URL will be provided.

* The container will appear in the left-hand list.
* Selecting an item will show its detailed information on the right.
* You can edit settings from the right panel.
* Clicking the "Dashboard" button will redirect you to a detailed management page with container status and activity logs.

**[API Request]**

Once the container status is RUNNING, you can access the AI inference API using the exposed endpoint.  
Replace the host part with your container's Endpoint URL.

**Sample API Request:**

curl --request POST \

--url ${ENDPOINT\_URL}/api/v1/chat/completions \

--header "Accept: application/json" \

--header "Authorization: Bearer ${YOUR\_API\_KEY}" \

--header "Content-Type: application/json" \

--data '

{

"messages": [

{

"role": "system",

"content": "You are a helpful assistant."

},

{

"role": "user",

"content": "Write a haiku about recursion in programming."

}

]

}'

Air Container enables you to deploy containerized AI services quickly and efficiently. Here's how it works:

스크린샷, 멀티미디어 소프트웨어, 그래픽 소프트웨어, 소프트웨어이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

1. **Select a Project**: Go to your Project Dashboard and choose the organization and project you created during onboarding.

스크린샷, 멀티미디어 소프트웨어, 소프트웨어, 그래픽 소프트웨어이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

1. **Choose Service Type**: Select "Container Service."

멀티미디어 소프트웨어, 스크린샷이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

1. **Enter General Settings**: Input your container name, select cloud type (Air Cloud or Air Cloud+), and define the number of replicas.

스크린샷, 텍스트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

1. **Container Image Settings**: Enter the image path (e.g., Docker Hub, AWS ECR), your credentials, and specify registry settings.

텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

1. **Resource Settings**:
   * **General Mode**: Define instance type (e.g., RTX 4070, 4090).
   * **Autoscaling Mode**: Define minimum and maximum replicas (1 to 30). Higher values require manual approval. For values above this range, please contact AIEEV.

스크린샷, 텍스트, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

1. **Advanced Settings**: Configure model start commands, port, health check, and environment variables.

스크린샷, 텍스트, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

1. **Review & Deploy**: Review all settings and launch your container. Images are cached after the first pull, reducing startup time.

**4. Monitoring & API Testing**

* **Single Request Test**: Use Postman or similar tools to test your vLLM container endpoint.
* **Autoscaling Test**: Use Locust to simulate load and validate replica scaling.
* **Usage Metrics**: Monitor replica count, request rate, and system response time.
* **Runtime Logs**:
  + Logs are displayed per container instance and replica.
  + You can access real-time and historical logs from the Air Cloud dashboard.
  + Logs are shown in reverse chronological order.
  + Use filters to view logs by time range, container, or instance ID.
  + Logs include startup command output, health check status, error messages, and stdout/stderr of model servers.
  + If your container fails, logs will be preserved for a limited retention window for debugging.
* **Settings**: You can update endpoint settings only when the container is stopped.

**5. Troubleshooting**

Below are common issues and how to resolve them based on real user deployments and our internal diagnostics:

**[Common HTTP Errors]**

* **401 Unauthorized**: Authentication failed. Please verify that your API key or access credentials are correct.
* **404 Not Found**: The requested container ID or endpoint path is invalid. Double-check the container ID or route.
* **500 Internal Server Error**: An internal server error has occurred. Try again later or contact the Air Cloud support team.

Below are common issues and how to resolve them based on real user deployments and our internal diagnostics:

**[A]. API Key Authentication Failure**

* **Symptom**: Fails to authenticate API request
* **Cause**: Expired or incorrect API key
* **Fix**:
  + Ensure your key is still valid (check expiration)
  + Generate a new API key from the Air Cloud dashboard

**[B]. GPU Assignment Errors**

* **Symptom**: No GPU detected on container startup
* **Cause**: Node instability or driver mismatch
* **Fix**:
  + Re-deploy to a new instance
  + Check compatibility of driver with your container runtime

**[C]. Network Connectivity Failures**

* **Symptom**: Container cannot reach external endpoints
* **Cause**: Port configuration, DNS issues, or firewall blocks
* **Fix**:
  + Verify container exposes correct port and service is listening
  + Use curl inside container to test outbound connectivity
  + Ensure no security group restrictions block egress traffic

**[D]. Health Check Failures / Stuck in Initializing**

* **Symptom**: Deployment hangs on health check
* **Cause**: Container is not returning 200 OK on /health
* **Fix**:
  + Confirm that the correct health check path and port are exposed
  + Check logs for stack traces or model loading delays

**[E]. Autoscaling Not Triggering**

* **Symptom**: No scaling despite high load
* **Cause**: Scaling threshold too high, or unhealthy replicas
* **Fix**:
  + Tune autoscaling parameters (e.g., CPU threshold, latency window)
  + Verify all replicas pass health checks

**[F]. 502/503 Gateway Errors**

* **Symptom**: API responds with server error
* **Cause**: App not bound to expected port or crashed during startup
* **Fix**:
  + Verify service listens on 0.0.0.0:<PORT>
  + Increase health check timeout if model loading takes long

**6. Contact**

For additional support, contact:

* **Hyunkyung Kim**: [hk.kim@aieev.com](mailto:hk.kim@aieev.com) / +82-10-9888-4696
* **Byoungsoo Park**: [bspark@aieev.com](mailto:bspark@aieev.com) / +82-10-4102-4050