



ENKIDU

— MIOVSKI GROUP —

ARQUITECTURA CONCEPTUAL

Helios Grid

El plano de control de la inteligencia distribuida — desplegar modelos, vigilar la salud y responder a eventos en miles de nodos, elevando solo telemetría y señal, nunca datos brutos. Estructurado por Tangram, gobernado por Origi.

Español · Edición resumida

Basada en la versión 0.1 · 4 de julio de 2026

Preparado por MIOVSKI GROUP

Clasificación · No clasificado · Borrador para revisión interna

La edición completa en inglés prevalece

Acerca de esta edición

Este documento es una edición resumida, en español, de « Helios Grid — The control plane for distributed intelligence », versión 0.1 (4 de julio de 2026), preparada por Miovski Group. Cada sección del original está representada de forma abreviada. En caso de divergencia, la edición completa en inglés prevalece.

Clasificación: no clasificado — borrador para revisión interna.

Resumen

Un tejido de inteligencia distribuida de la clase Northwind triunfa o fracasa en un problema que no es ni la detección ni la inferencia: operar la flota. De cientos a miles de nodos de borde — portados por equipos, montados en vehículos, integrados en plataformas autónomas o fijados a postes — deben ser enrolados, configurados, mantenidos sanos, actualizados con nuevos modelos y políticas, y recuperados cuando fallan, a través de enlaces intermitentes, disputados y pobres en ancho de banda. Hacerlo a mano no escala más allá de un puñado de nodos; hacerlo con un servicio convencional de gestión de dispositivos en la nube viola las restricciones de soberanía, clasificación y desconexión bajo las que viven estos tejidos.

Helios Grid es la respuesta de Enkidu: un plano de control soberano y autoalojado para la inteligencia distribuida, y el tercer marco fundamental de la pila Enkidu. Donde Tangram da al tejido su estructura operativa (zonas, continuos, traspasos contractuales) y Origi su confianza en ejecución (aplicación de políticas, zonas de operación de confianza, cadena de custodia), Helios le da su operabilidad: ciclo de vida de la flota, despliegue escalonado de modelos y software, observabilidad de la salud y respuesta automatizada a eventos. Su disciplina definitoria es el principio « señal, no datos »: Helios solo transmite hacia arriba telemetría y señal, nunca datos brutos — visibilidad a escala de flota sin ancho de banda a escala de flota, y sin crear jamás un segundo camino por el que los datos de misión pudieran salir de un nodo.

Helios sigue la premisa establecida en los marcos Enkidu — determinismo donde importa, razonamiento donde ayuda — y añade su corolario: declaración donde escala. El estado de la flota se declara, se firma y se reconcilia por tracción (pull), de modo que un nodo que estuvo a oscuras un día converge al mismo estado que un nodo que nunca perdió su enlace. La arquitectura adopta deliberadamente los patrones más probados del ecosistema NVIDIA de IA en el borde — la orquestación de flotas de Fleet Command, los servicios de base de Jetson Platform Services, la actualización a prueba de fallos A/B, la gestión de modelos en caliente de Triton, la telemetría GPU de clase DCGM, la recuperación autónoma de Mission Control, la reconciliación GitOps por tracción — y los refunde tras un único plano de control gobernado y soberano, supervisado por Origi y estructurado por Tangram.

1. Propósito y alcance

Esto es una arquitectura conceptual, no un manual de implementación: define la estructura de Helios Grid, sus decisiones estructurales y sus contratos con los demás marcos Enkidu. El despliegue de referencia es Northwind — la versión más difícil del problema: cuatro niveles de cómputo, enlaces disputados, clasificación de seguridad, anfitriones autónomos. La misma arquitectura, con enlaces civiles, es el plano de control de Lamassu (detección urbana) y de despliegues fijos como las unidades de salud inteligentes. Un solo plano de control, muchos tejidos.

Lo que Helios no es: ni una plataforma de datos, ni un motor de fusión, ni un marco de confianza. Nunca ingiere telemetría de sensores, nunca porta observaciones o evaluaciones, nunca emite un veredicto de misión y nunca amplía la autoridad de un agente. Su negativa a tocar los datos de misión no es una limitación; es el diseño.

- Ciclo de vida de la flota a escala — enrolamiento sin contacto, gemelos de nodo, estado deseado declarado, reconciliación por tracción.
- Distribución segura de artefactos — despliegues firmados, escalonados, con puertas de salud, instalación A/B y reversión automática.
- Observabilidad a escala de flota — visibilidad de salud y rendimiento en cada nivel, dentro de una envolvente estricta de telemetría.
- Respuesta autónoma a eventos — detección, clasificación y remediación automatizada acotada, con escalada humana más allá del runbook.

2. Premisa de diseño y principios

Helios descansa en una frase: declarar en el centro, reconciliar localmente, informar por excepción. El centro declara lo que la flota debe ser; cada nodo tira hacia esa declaración a su propio ritmo y sobrevive a cualquier interrupción; y la flota solo eleva la señal que un operador necesita — salud, desviación y eventos — nunca los datos que la misión produce.

- 01 · El control nunca porta datos — ninguna carga de misión transita por un canal Helios; los esquemas de telemetría no pueden expresar estructuralmente contenido de misión.
- 02 · El estado se declara y se tira — el estado deseado se firma en el centro; los nodos se reconcilian por tracción; la desconexión es un estado normal, no un fallo.
- 03 · Todo artefacto se firma y se admite — Helios distribuye; Origi admite. Una firma inválida o una procedencia rota detiene la instalación en el nodo, ordene lo que ordene el centro.
- 04 · Las actualizaciones son a prueba de fallos por construcción — toda instalación mutable usa un mecanismo A/B con reversión automática; un fallo degrada al estado anterior, nunca a un nodo inservible.
- 05 · La remediación está acotada — las respuestas automatizadas provienen de un runbook firmado de acciones preaprobadas con límites explícitos de radio de impacto; lo demás escala a un humano.
- 06 · El propio plano de control está gobernado — el Agent, el Warden y el Core son agentes bajo Origi: zonas de confianza, política de denegación por defecto, cadena de custodia. El operador de la flota rinde cuentas igual que la flota.

¿Por qué no un servicio de gestión de dispositivos en la nube? Helios adopta lo que ese patrón demostró — una consola única, aprovisionamiento de un toque, gestión remota del ciclo de vida — pero rechaza su modelo de despliegue: los tejidos Enkidu operan bajo restricciones de soberanía de datos, clasificación (Protected B en Canadá) y operaciones desconectadas que un plano de control en una nube de terceros no puede satisfacer. Helios se autoaloja, por tanto, en el nivel central del propio tejido, hereda su frontera de acreditación y funciona con la WAN totalmente ausente.

3. Posición en la pila Enkidu

- Tangram — ¿Dónde ocurre el trabajo, y bajo qué condiciones de frontera? — Estructura.
- Origi — ¿Puede ocurrir esta acción, y podemos probar lo que ocurrió? — Confianza.
- Helios — ¿Está el propio tejido sano, actualizado y recuperable? — Operaciones.

Los tres se entrelazan sin solaparse: Helios está estructurado por Tangram (los segmentos de flota son zonas; una ola de despliegue es una operación de zona con condiciones de entrada y salida; un despliegue fallido entra en una zona de error diseñada); Helios está gobernado por Origi (distribuir un artefacto es una propuesta; la admisión de Origi — firma, procedencia, vínculo de política — es el commit); y tanto Origi como Tangram son operados por Helios (las políticas y las definiciones de zona son a su vez artefactos versionados). Ningún componente puede a la vez decidir y otorgar: Helios observa y ejecuta; Origi juzga; Tangram acota. Un plano de control comprometido puede, como mucho, proponer.

4. Topología de referencia — Helios sobre el tejido Northwind

- Nodo FIB (clase Orin) — Helios Agent — identidad criptográfica y enrolamiento; bucle de reconciliación; descarga, verificación e instalación A/B de artefactos; telemetría local; ejecutor local de runbook.
- Concentrador FIB (clase Thor) — Helios Warden — subplano de control de su grupo: caché de estado deseado y artefactos; agregación y submuestreo de telemetría; evaluación local de reglas de alerta; coordinación de olas; ganchos de recuperación fuera de banda.
- FOB-IFC (clase GB300) — Helios Core (avanzado) — registro y gemelos a escala de teatro; orquestación de despliegues sobre docenas de concentradores; lago de telemetría y motor de eventos; consola avanzada de flota.
- Central IFC (clase NVL72) — Helios Core (central) — registro de flota de flotas; almacén de estado deseado y registro de artefactos autoritativos; telemetría de largo plazo; integración con la gestión de infraestructura del centro de datos.

Dos propiedades dominan. Primera, el concentrador es la unidad de autonomía: un Warden retiene todo lo que su grupo necesita, de modo que un grupo aislado del FOB-IFC sigue reconciliándose, vigilándose y remediándose localmente. Segunda, la distribución sigue al tejido: un artefacto de modelo viaja una sola vez por enlace — una actualización de 2 GB a un grupo de 40 nodos cuesta al enlace de retaguardia un transferencia, no cuarenta.

5. Los tres planos de Helios

- Plano de estado — gemelos de nodo, almacén de estado deseado, registro de artefactos. Determinista, versionado, firmado: la única fuente de verdad.
- Plano de flujo — reconciliación, distribución, transporte de telemetría. Por tracción, reanudable, clasificado por prioridades, acotado por la envolvente — diseñado para el peor enlace de la flota, no el mejor.
- Plano de respuesta — motor de eventos, reglas de alerta, runbooks, rutas de escalada. Automatización acotada con una costura humana diseñada.

Una cuarta preocupación — la gobernanza — deliberadamente no es un plano de Helios: es Origi, que envuelve a los tres.

6. Gestión del ciclo de vida de la flota

La vida de un nodo empieza antes del despliegue: en el aprovisionamiento recibe su identidad criptográfica (arraigada en hardware donde el nivel lo permite), su perfil de nivel y una caché básica de estado deseado. En el campo, el primer arranque sigue el enrolamiento sin contacto; como el estado deseado y los modelos base viajan en el dispositivo, un nodo es útil antes de estar conectado y gestionado en cuanto lo está. El enrolamiento crea el gemelo de nodo — el registro autoritativo sobre el que razonan los operadores y el motor de eventos.

La reconciliación es idempotente y reanudable, según el modelo GitOps: un estado deseado declarado y versionado que agentes ligeros tiran y convergen — el único modelo que sobrevive a los perfiles de despliegue de Northwind, incluida la operación MANET con infraestructura denegada. Los nodos dejan la flota por tres vías, todas diseñadas: retiro ordenado (claves revocadas, gemelo archivado, artefactos borrados), compromiso sospechado (cuarentena de Origi, enrolamiento revocado, la malla instruida para rehuirlo, evidencia preservada) y pérdida física (el cifrado en reposo y la custodia de claves acotan lo que un nodo capturado puede revelar). Una flota que solo planifica añadir nodos es media flota.

7. Distribución de artefactos y modelos

Todo lo que un nodo ejecuta u obedece es un artefacto: modelos especialistas afinados y cuantizados (y sus adaptadores), imágenes de contenedores, definiciones de zonas Tangram, paquetes de políticas Origi, conectores de ingesta, imágenes de sistema base. El registro de artefactos del Core central es el análogo soberano de un registro privado: un catálogo firmado del que puede extraerse todo objeto desplegable — y nada más. Un modelo que no puede mostrar su procedencia no entra al registro; un modelo fuera del registro no llega a un nodo.

Ningún artefacto alcanza toda la flota de una vez: un despliegue declara sus olas — canario, grupo, teatro — y las puertas de salud entre ellas; el fallo de una puerta congela la ola y dispara la reversión automática de los nodos ya actualizados. En la instalación todo es transaccional: estrategia A/B completa para artefactos de sistema (la ranura que falla el control de salud nunca se consolida); para los modelos, la API de control del servidor de inferencia permite cargar, descargar y reemplazar en caliente, sin pérdida de disponibilidad. Las transferencias delta mantienen viables los enlaces restringidos; los concentradores portan además capacidad de recuperación fuera de banda — ciclado de alimentación por hardware y flasheo en modo recuperación — para revivir un nodo sin visita al sitio.

8. Telemetría, salud y el principio « señal, no datos »

La observabilidad de Helios se construye hacia atrás desde su restricción de ancho de banda y su restricción de confianza, que coinciden: el borde reduce; el tejido reenvía señal; los datos brutos nunca abandonan el nodo por un canal de control. Los esquemas de telemetría no tienen campo alguno en el que pudiera viajar un fotograma de vídeo, una captura RF o una evaluación.

Cada nodo y concentrador posee una envolvente de telemetría firmada: presupuesto por intervalo, clases de prioridad (eventos > puntuaciones de salud > series submuestreadas > registros bajo demanda) y comportamiento degradado. Cada enlace ascendente porta un orden de magnitud menos que el inferior. El Warden calcula una puntuación de salud por nodo — un número con el que clasificar mil nodos — mientras las series completas siguen consultables en el nivel que las posee; las reglas de alerta se evalúan en el nivel más bajo que tiene el dato, de modo que un nodo hace sonar su propia alarma incluso solo y a oscuras. La salud de los modelos es una clase de métricas de primer orden: distribuciones de confianza, tasas de abstención, indicadores de deriva — el alimento operativo de la vigilancia de deriva de Origi y el disparador de las peticiones de reentrenamiento.

9. Respuesta a eventos y remediación automatizada

Los eventos — alertas, fallos de puerta, anomalías de enrolamiento, degradaciones de confianza de Origi, cambios de estado de enlace — fluyen al motor de eventos de cada Core, se clasifican por gravedad y radio de impacto, se correlacionan a través de la flota (un artefacto defectuoso parece cincuenta eventos simultáneos; el motor debe ver una sola causa) y se contrastan con el runbook. Una entrada de runbook es una respuesta automatizada firmada, versionada y preaprobada, con precondiciones y límites explícitos. La autonomía es deliberadamente conservadora — reiniciar, revertir, aislar, descargar, pedir recuperación — y asimétrica en el mismo sentido que las zonas de confianza de Origi: una acción automatizada puede reducir la actividad o capacidad de un nodo al instante, pero nada automatizado aumenta jamás la capacidad o el alcance. Todo lo que sale del runbook escala, con el gemelo, la cadena de eventos correlacionada y la telemetría pertinente ya reunidas en la vista del operador.

10. Seguridad y gobernanza — Helios bajo Origi

Helios es el objetivo de mayor valor de cualquier tejido que gestione: toca cada nodo y distribuye artefactos ejecutables. Su postura de seguridad no es, por tanto, autoafirmada; es la de Origi, aplicada a Helios como a cualquier otro sistema de agentes: cada enlace autenticado mutuamente y cifrado; autoridad de denegación por defecto (ningún componente puede ampliar su propia política); la admisión por encima de la distribución (la verificación corre en el nodo, en el último momento antes de instalar — una caché comprometida puede, como mucho, entregar un artefacto que el nodo rechazará); separación de funciones (evaluación, juicio y ejecución son componentes distintos que consumen las salidas firmadas de los otros); auditabilidad total (« ¿quién puso este modelo en este nodo, cuándo, bajo qué autoridad, y qué reemplazó? » siempre se responde desde la traza); y disciplina de clasificación sobre el propio tráfico de control.

11. Alineación con las herramientas NVIDIA

Helios no reinventa lo que el ecosistema NVIDIA de IA en el borde ya demostró; adopta, endurece y autoaloja: el patrón operativo de Fleet Command (rechazando su modelo en la nube), los servicios de Jetson Platform Services como sustrato de telemetría en el dispositivo, la estrategia A/B validada Mender/L4T como invariante de instalación, la recuperación fuera de banda de clase Allxon, el servicio de modelos Triton en modo de control explícito, el pipeline de producción de modelos de patrón TAO publicando exclusivamente en el registro Helios, la telemetría DCGM en los niveles de servidor, Mission Control como plano de gestión interno del Central IFC, y la reconciliación GitOps con una distribución ligera de Kubernetes de clase K3s en los niveles concentrador y central.

12. Escalabilidad y comportamiento ante fallos

El modelo de escala sigue al tejido: docenas de nodos por Warden, docenas de Wardens por Core avanzado, Cores federados en el centro. Bajo pérdida: un nodo aislado sigue sintiendo e infiriendo, y se reconcilia al volver el enlace; un grupo aislado del centro sigue operando bajo su Warden; un Core avanzado perdido deja a los grupos autónomos hasta reconstruirse desde el Core central. La prueba de diseño: ningún fallo del plano de control debe degradar jamás el plano de datos — perder Helios significa perder la gestionabilidad, nunca la misión.

13. Madurez y camino a seguir

La entrega se alinea con las fases de Northwind: primero el Agent, el Warden y el Core avanzado con enrolamiento, reconciliación y telemetría; después la orquestación completa de olas, la recuperación fuera de banda y el motor de eventos; por último el Core central federado, la integración con Mission Control y la distribución asistida por pares a través de la malla. Las cuestiones abiertas quedan registradas como compromisos para el diseño detallado.

Conclusión

Tangram hizo legibles los tejidos Enkidu. Origi los hizo defendibles. Helios Grid los hace operables — y a escala de flota, la operabilidad no es una comodidad sino una precondition: mil nodos de borde sin gobernar, sin parchear, sin observar no son una capacidad, sino un pasivo con antenas.

La contribución de Helios es un plano de control lo bastante disciplinado como para confiarle la flota entera: estado declarado y tirado en vez de empujado y esperado; artefactos firmados, escalonados, con puertas y reversibles; visibilidad construida con señal y no con datos, para ver la flota sin filtrarla; remediación automatizada solo dentro de límites firmados, con la costura humana diseñada; y el plano entero sometido al juicio de Origi y dentro de la estructura de Tangram, de modo que el operador de la flota responda al mismo estándar que la flota.

Declarar en el centro. Reconciliar localmente. Informar por excepción. Y no dejar nunca que el plano de control toque los datos.