



E N K I D U

— M I O V S K I G R O U P —

A R C H I T E C T U R E C O N C E P T U E L L E

Helios Grid

Le plan de contrôle de l'intelligence distribuée — déployer les modèles, surveiller la santé et répondre aux événements sur des milliers de nœuds, en ne faisant remonter que la télémétrie et le signal, jamais les données brutes. Structuré par Tangram, gouverné par Origi.

Français · Édition condensée

D'après la version 0.1 · 4 juillet 2026

Préparé par MIOVSKI GROUP

Classification · Non classifié · Ébauche pour examen interne

L'édition anglaise complète fait foi

À propos de cette édition

Ce document est une édition condensée, en français, de « Helios Grid — The control plane for distributed intelligence », version 0.1 (4 juillet 2026), préparée par Miovski Group. Chaque section de l'original y est représentée sous forme abrégée. En cas de divergence, l'édition anglaise complète fait foi. Classification : non classifié — ébauche pour examen interne.

Résumé

Un tissu d'intelligence distribuée de la classe Northwind réussit ou échoue sur un problème qui n'est ni la détection ni l'inférence : l'exploitation de la flotte. Des centaines à des milliers de nœuds de périphérie — portés par des équipes, montés sur des véhicules, intégrés à des plateformes autonomes ou fixés à des poteaux — doivent être enrôlés, configurés, maintenus en santé, mis à jour avec de nouveaux modèles et politiques, et récupérés en cas de panne, à travers des liaisons intermittentes, contestées et pauvres en bande passante. Le faire à la main ne passe pas l'échelle au-delà d'une poignée de nœuds ; le faire avec un service infonuagique classique de gestion d'appareils viole les contraintes de souveraineté, de classification et de déconnexion sous lesquelles vivent ces tissus.

Helios Grid est la réponse d'Enkidu : un plan de contrôle souverain et auto-hébergé pour l'intelligence distribuée, et le troisième cadre fondamental de la pile Enkidu. Là où Tangram donne au tissu sa structure opérationnelle (zones, continuums, transferts contractuels) et Origi sa confiance à l'exécution (application des politiques, zones d'opération de confiance, chaîne de possession), Helios lui donne son exploitabilité : cycle de vie de la flotte, déploiement échelonné des modèles et logiciels, observabilité de la santé et réponse automatisée aux événements. Sa discipline définitoire est le principe « du signal, pas des données » : Helios ne fait remonter que la télémétrie et le signal, jamais de données brutes — une visibilité à l'échelle de la flotte sans la bande passante de la flotte, et sans jamais créer un second chemin par lequel des données de mission pourraient quitter un nœud.

Helios suit la prémisse établie à travers les cadres Enkidu — le déterminisme là où il compte, le raisonnement là où il aide — et y ajoute son corollaire : la déclaration là où elle passe l'échelle. L'état de la flotte est déclaré, signé et réconcilié par traction (pull), de sorte qu'un nœud resté dans le noir une journée converge vers le même état qu'un nœud qui n'a jamais perdu sa liaison. L'architecture adopte délibérément les patrons les plus éprouvés de l'écosystème NVIDIA d'IA en périphérie — l'orchestration de flotte de Fleet Command, les services de base embarqués de Jetson Platform Services, la mise à jour à sécurité intégrée A/B, la gestion de modèles à chaud de Triton, la télémétrie GPU de classe DCGM, la récupération autonome de Mission Control, la réconciliation GitOps par traction — et les refond derrière un unique plan de contrôle gouverné et souverain, supervisé par Origi et structuré par Tangram.

1. Objet et portée

Ceci est une architecture conceptuelle, non un manuel d'implémentation : elle définit la structure d'Helios Grid, ses décisions porteuses et ses contrats avec les autres cadres Enkidu. Le déploiement de référence est Northwind — la version la plus difficile du problème : quatre niveaux de calcul, liaisons contestées, classification de sécurité, hôtes autonomes. La même architecture, avec des liaisons civiles, est le plan de contrôle de Lamassu (détection urbaine) et des déploiements fixes comme les unités de santé intelligentes. Un seul plan de contrôle, plusieurs tissus.

Ce qu'Helios n'est pas : ni une plateforme de données, ni un moteur de fusion, ni un cadre de confiance. Il n'ingère jamais de télémétrie de capteurs, ne porte jamais d'observations ou d'évaluations, ne rend jamais de verdict de mission et n'élargit jamais l'autorité d'un agent. Son refus de toucher aux données de mission n'est pas une limite ; c'est le dessein.

- Cycle de vie de la flotte à l'échelle — enrôlement sans contact, jumeaux de nœud, état désiré déclaré, réconciliation par traction.
- Distribution sûre des artefacts — déploiements signés, échelonnés, à barrières de santé, avec installation A/B et retour arrière automatique.
- Observabilité à l'échelle de la flotte — visibilité de santé et de performance à chaque niveau, dans une enveloppe de télémétrie stricte.
- Réponse autonome aux événements — détection, classification et remédiation automatisée bornée, avec escalade humaine au-delà du runbook.

2. Prémisse de conception et principes

Helios repose sur une phrase : déclarer au centre, réconcilier localement, rapporter par exception. Le centre déclare ce que la flotte doit être ; chaque nœud tire vers cette déclaration à son propre rythme et survit à toute interruption ; et la flotte ne fait remonter que le signal dont un opérateur a besoin — santé, écart et événements — jamais les données que la mission produit.

- 01 · Le contrôle ne porte jamais de données — aucune charge utile de mission ne transite par un canal Helios ; les schémas de télémétrie ne peuvent structurellement pas exprimer de contenu de mission.
- 02 · L'état est déclaré et tiré — l'état désiré est signé au centre ; les nœuds se réconcilient par traction ; la déconnexion est un état normal, pas une panne.
- 03 · Chaque artefact est signé et admis — Helios distribue ; Origi admet. Une signature invalide ou une provenance brisée arrête l'installation au nœud, quoi qu'ait ordonné le centre.
- 04 · Les mises à jour sont à sécurité intégrée par construction — chaque installation mutable utilise un mécanisme A/B avec retour arrière automatique ; un échec dégrade vers l'état antérieur, jamais vers un nœud brique.
- 05 · La remédiation est bornée — les réponses automatisées proviennent d'un runbook signé d'actions pré-approuvées avec des limites explicites de rayon d'impact ; le reste escalade vers un humain.
- 06 · Le plan de contrôle est lui-même gouverné — l'Agent, le Warden et le Core sont des agents sous Origi : zones de confiance, politique de refus par défaut, chaîne de possession. L'opérateur de la flotte est aussi redevable que la flotte.

Pourquoi pas un service infonuagique de gestion d'appareils ? Helios adopte ce que ce patron a prouvé — une console unique, l'approvisionnement en un geste, la gestion du cycle de vie à distance — mais rejette son modèle de déploiement : les tissus Enkidu opèrent sous des contraintes de souveraineté des données, de classification (Protégé B au Canada) et d'opérations déconnectées qu'un plan de contrôle infonuagique tiers ne peut satisfaire. Helios est donc auto-hébergé au niveau central du tissu lui-même, hérite de sa frontière d'accréditation et fonctionne avec le WAN entièrement absent.

3. Position dans la pile Enkidu

- Tangram — Où le travail se produit-il, et sous quelles conditions aux frontières ? — Structure.
- Origi — Cette action peut-elle se produire, et pouvons-nous prouver ce qui s'est produit ? — Confiance.
- Helios — Le tissu lui-même est-il en santé, à jour et récupérable ? — Opérations.

Les trois s'emboîtent sans se chevaucher : Helios est structuré par Tangram (les segments de flotte sont des zones ; une vague de déploiement est une opération de zone avec conditions d'entrée et de sortie ; un déploiement raté entre dans une zone d'erreur conçue) ; Helios est gouverné par Origi (distribuer un artefact est une proposition ; l'admission d'Origi — signature, provenance, liaison de politique — est le commit) ; et Origi comme Tangram sont exploités par Helios (les politiques et les définitions de zones sont elles-mêmes des artefacts versionnés). Nul composant ne peut à la fois décider et accorder : Helios observe et exécute ; Origi juge ; Tangram borne. Un plan de contrôle compromis peut, au pire, proposer.

4. Topologie de référence — Helios sur le tissu Northwind

- Nœud FIB (classe Orin) — Agent Helios — identité cryptographique et enrôlement ; boucle de réconciliation ; téléchargement, vérification et installation A/B des artefacts ; télémétrie locale ; exécuter local de runbook.
- Concentrateur FIB (classe Thor) — Warden Helios — sous-plan de contrôle de sa grappe : cache d'état désiré et d'artefacts ; agrégation et sous-échantillonnage de la télémétrie ; évaluation locale des règles d'alerte ; coordination des vagues ; crochets de récupération hors bande.
- FOB-IFC (classe GB300) — Core Helios (avancé) — registre et jumeaux à l'échelle du théâtre ; orchestration des déploiements sur des dizaines de concentrateurs ; lac de télémétrie et moteur d'événements ; console avancée de flotte.
- Central IFC (classe NVL72) — Core Helios (central) — registre de flotte de flottes ; magasin d'état désiré et registre d'artefacts faisant autorité ; télémétrie de long terme ; intégration à la gestion d'infrastructure du centre de données.

Deux propriétés dominant. D'abord, le concentrateur est l'unité d'autonomie : un Warden détient tout ce dont sa grappe a besoin, si bien qu'une grappe coupée du FOB-IFC continue de se réconcilier, de surveiller et de remédier localement. Ensuite, la distribution suit le tissu : un artefact de modèle voyage une seule fois par liaison — une mise à jour de 2 Go vers une grappe de 40 nœuds coûte un transfert à la liaison arrière, pas quarante.

5. Les trois plans d'Helios

- Plan d'état — jumeaux de nœud, magasin d'état désiré, registre d'artefacts. Déterministe, versionné, signé : la source unique de vérité.
- Plan de flux — réconciliation, distribution, transport de télémétrie. Par traction, reprenable, classé par priorités, borné par l'enveloppe — conçu pour la pire liaison de la flotte, pas la meilleure.
- Plan de réponse — moteur d'événements, règles d'alerte, runbooks, chemins d'escalade. Automatisation bornée avec une couture humaine conçue.

Une quatrième préoccupation — la gouvernance — n'est délibérément pas un plan d'Helios : c'est Origi, qui enveloppe les trois.

6. Gestion du cycle de vie de la flotte

La vie d'un nœud commence avant le déploiement : à l'approvisionnement, l'appareil reçoit son identité cryptographique (enracinée dans le matériel où le niveau le permet), son profil de niveau et un cache d'état désiré de base. Sur le terrain, le premier démarrage suit l'enrôlement sans contact ; parce que l'état désiré et les modèles de base voyagent sur l'appareil, un nœud est utile avant d'être connecté et géré dès qu'il l'est. L'enrôlement crée le jumeau de nœud — l'enregistrement faisant autorité sur lequel raisonnent les opérateurs et le moteur d'événements.

La réconciliation est idempotente et reprenable, sur le modèle GitOps : un état désiré déclaré et versionné que des agents légers tirent et convergent — le seul modèle qui survive aux profils de déploiement de Northwind, y compris l'opération MANET en infrastructure refusée. Les nœuds quittent la flotte par trois voies, toutes conçues : retraite en règle (clés révoquées, jumeau archivé, artefacts effacés), compromission soupçonnée (quarantaine Origi, enrôlement révoqué, maillage instruit de l'éviter, preuves préservées) et perte physique (chiffrement au repos et garde des clés bornent ce qu'un nœud capturé peut divulguer). Une flotte qui ne planifie que l'ajout de nœuds est une demi-flotte.

7. Distribution des artefacts et des modèles

Tout ce qu'un nœud exécute ou obéit est un artefact : modèles spécialistes affinés et quantifiés (et leurs adaptateurs), images de conteneurs, définitions de zones Tangram, lots de politiques Origi, connecteurs d'ingestion, images de base système. Le registre d'artefacts au Core central est l'analogue souverain d'un registre privé : un catalogue signé duquel tout objet déployable — et rien d'autre — peut être tiré. Un modèle qui ne peut montrer sa provenance ne peut entrer au registre ; un modèle hors du registre ne peut atteindre un nœud.

Aucun artefact n'atteint toute la flotte d'un coup : un déploiement déclare ses vagues — canari, grappe, théâtre — et les barrières de santé entre elles ; l'échec d'une barrière gèle la vague et déclenche le retour arrière automatique des nœuds déjà mis à jour. À l'installation, tout est transactionnel : stratégie A/B complète pour les artefacts système (l'emplacement qui échoue au contrôle de santé ne s'engage jamais) ; pour les modèles, l'API de contrôle du serveur d'inférence permet le chargement, le déchargement et le remplacement à chaud, sans perte de disponibilité. Les transferts delta gardent les liaisons contraintes viables ; les concentrateurs portent en outre une capacité de récupération hors bande — cycle d'alimentation matériel et re-flashage en mode récupération — pour ranimer un nœud sans visite sur site.

8. Télémétrie, santé et le principe « du signal, pas des données »

L'observabilité d'Helios est construite à rebours depuis sa contrainte de bande passante et sa contrainte de confiance, qui se trouvent coïncider : la périphérie réduit ; le tissu fait suivre du signal ; les données brutes ne quittent jamais le nœud sur un canal de contrôle. Les schémas de télémétrie n'ont aucun champ où une trame vidéo, une capture RF ou une évaluation pourrait voyager.

Chaque nœud et concentrateur détient une enveloppe de télémétrie signée : budget par intervalle, classes de priorité (événements > scores de santé > séries sous-échantillonnées > journaux sur demande) et comportement dégradé. Chaque liaison montante porte un ordre de grandeur de moins que celle du dessous. Le Warden calcule un score de santé par nœud — un nombre pour classer mille nœuds — tandis que les séries complètes restent interrogeables au niveau qui les détient ; les règles d'alerte s'évaluent au niveau le plus bas qui possède la donnée, si bien qu'un nœud sonne sa propre alarme même seul dans le noir. La santé des modèles est une classe de métriques de premier rang : distributions de confiance, taux d'abstention, indicateurs de dérive — l'alimentation opérationnelle de la surveillance de dérive d'Origi et le déclencheur des demandes de ré-entraînement.

9. Réponse aux événements et remédiation automatisée

Les événements — alertes, échecs de barrière, anomalies d'enrôlement, rétrogradations de confiance Origi, changements d'état de liaison — affluent vers le moteur d'événements de chaque Core, sont classés par gravité et rayon d'impact, corrélés à travers la flotte (un mauvais artefact ressemble à cinquante événements simultanés ; le moteur doit voir une seule cause) et confrontés au runbook. Une entrée de runbook est une réponse automatisée signée, versionnée et pré-approuvée, aux préconditions et limites explicites.

L'autonomie est délibérément conservatrice — redémarrer, revenir en arrière, isoler, délester, demander une récupération — et asymétrique dans le même sens que les zones de confiance d'Origi : une action automatisée peut réduire l'activité ou la capacité d'un nœud instantanément, mais rien d'automatisé n'augmente jamais la capacité ou la portée. Tout ce qui sort du runbook escalade, avec le jumeau, la chaîne d'événements corrélée et la télémétrie pertinente déjà assemblées dans la vue de l'opérateur.

10. Sécurité et gouvernance — Helios sous Origi

Helios est la cible de plus haute valeur de tout tissu qu'il gère : il touche chaque nœud et distribue des artefacts exécutables. Sa posture de sécurité n'est donc pas auto-affirmée ; c'est celle d'Origine, appliquée à Helios comme à tout autre système d'agents : chaque liaison mutuellement authentifiée et chiffrée ; autorité de refus par défaut (nul composant ne peut élargir sa propre politique) ; l'admission au-dessus de la distribution (la vérification s'exécute au nœud, au dernier moment avant l'installation — un cache compromis peut au pire livrer un artefact que le nœud refusera) ; séparation des devoirs (évaluation, jugement et exécution sont des composants distincts qui consomment leurs sorties signées respectives) ; auditabilité totale (« qui a mis ce modèle sur ce nœud, quand, sous quelle autorité, et qu'a-t-il remplacé ? » se répond toujours depuis la trace) ; et discipline de classification sur le trafic de contrôle lui-même.

11. Alignement sur l'outillage NVIDIA

Helios ne réinvente pas ce que l'écosystème NVIDIA d'IA en périphérie a déjà prouvé ; il adopte, durcit et auto-héberge : le patron opérationnel de Fleet Command (en rejetant son modèle infonuagique), les services de base Jetson Platform Services comme substrat de télémétrie embarquée, la stratégie A/B validée Mender/L4T comme invariant d'installation, la récupération hors bande de classe Allxon, le service de modèles Triton en mode contrôle explicite, le pipeline de production de modèles de patron TAO publiant exclusivement dans le registre Helios, la télémétrie DCGM aux niveaux serveurs, Mission Control comme plan de gestion interne du Central IFC, et la réconciliation GitOps avec une distribution Kubernetes légère de classe K3s aux niveaux concentrateur et central.

12. Passage à l'échelle et comportement en panne

Le modèle d'échelle suit le tissu : des dizaines de nœuds par Warden, des dizaines de Wardens par Core avancé, des Cores fédérés au centre. Sous perte : un nœud isolé continue de sentir et d'inférer, puis se réconcilie au retour de la liaison ; une grappe coupée du centre continue de fonctionner sous son Warden ; un Core avancé perdu laisse les grappes autonomes jusqu'à reconstruction depuis le Core central. Le test de conception : aucune panne du plan de contrôle ne doit jamais dégrader le plan de données — perdre Helios signifie perdre la gérabilité, jamais la mission.

13. Maturité et chemin à venir

La livraison s'aligne sur les phases de Northwind : d'abord l'Agent, le Warden et le Core avancé avec enrôlement, réconciliation et télémétrie ; puis l'orchestration complète des vagues, la récupération hors bande et le moteur d'événements ; enfin le Core central fédéré, l'intégration Mission Control et la distribution assistée par les pairs à travers le maillage. Les questions ouvertes sont consignées comme engagements pour la conception détaillée.

Conclusion

Tangram a rendu les tissus Enkidu lisibles. Origi les a rendus défendables. Helios Grid les rend exploitables — et à l'échelle d'une flotte, l'exploitabilité n'est pas une commodité mais une précondition : mille nœuds de périphérie non gouvernés, non corrigés, non observés ne sont pas une capacité, mais un passif muni d'antennes.

La contribution d'Helios est un plan de contrôle assez discipliné pour qu'on lui confie la flotte entière : un état déclaré et tiré plutôt que poussé et espéré ; des artefacts signés, échelonnés, à barrières et réversibles ; une visibilité bâtie sur le signal plutôt que sur les données, pour voir la flotte sans la faire fuir ; une remédiation automatisée seulement dans des bornes signées, avec la couture humaine conçue ; et le plan tout entier se tenant lui-même sous le jugement d'Origi et dans la structure de Tangram, de sorte que l'opérateur de la flotte soit tenu au même standard que la flotte.

Déclarer au centre. Réconcilier localement. Rapporter par exception. Et ne jamais laisser le plan de contrôle toucher aux données.