



Veille Technologique Hebdomadaire

Technique

2 octobre 2026 – Fréquence : Hebdomadaire (Vendredi)

Synthèse & Top 5 de la semaine

Inférence hétérogène (AMD ROCm), GitOps à l'échelle, guardrails de plateforme et agents d'entreprise : le socle d'ingénierie se stabilise.

Semaine 40 : Red Hat AI Inference 3.5 s'exécute sur GPU AMD en Podman rootless ; la communauté Argo ouvre la vision Argo CD 4.0 (ArgoCon NA) ; la policy Kubernetes se pense comme guardrail et non comme portail (Kyverno) ; OpenClaw Enterprise ouvre un plan de contrôle pour agents ; l'IA appliquée à la CI reste encadrée par du code déterministe.

1. Périmètre

Famille de veille : 2. Métiers, Compétences & Technologies (Indicateur Qualiopi 24).

Fenêtre de rétrospection : 26 septembre → 2 octobre 2026.

Fréquence d'exécution : Hebdomadaire (Vendredi).

Diffusion : [Radar Web Aperta](#) — [LinkedIn Aperta Scientia](#) — [X @ApertaScientia](#).

2. Sujet de la Semaine — Article Publié sur le Radar Web

Top 5 de la veille technologique (AS200 / AS300)

1. AMD GPU dans l'inférence LLM d'entreprise — Red Hat AI Inference 3.5 (ROCm) + Podman rootless Benchmark pas-à-pas d'un serveur d'inférence compatible OpenAI sur GPU AMD Radeon (ROCm), servi par **Red Hat AI Inference** (stack vLLM + llm-d) dans un conteneur **Podman rootless**, instrumenté par **GuideLLM** (méthodologie PSAP, paliers de concurrence de 1 à 650). L'accès GPU repose sur `--device=/dev/kfd --device=/dev/dri --group-add=keep-groups` et requiert le runtime OCI **crun** pour préserver les groupes ; l'image est [registry.redhat.io/rhail/vllm-rocm-rhel9:3.5.0](#). Conclusion : l'inférence d'entreprise n'est plus mono-constructeur, AMD devient un chemin de production supporté.

Source : [Red Hat Emerging Technologies — Benchmarking enterprise LLMs on AMD GPUs with Red Hat AI Inference and Podman](#) (28 septembre 2026).

2. CI augmentée par l'IA — Pipeline Failure Analyzer : l'IA sur 1 étape sur 5 Retour d'expérience Red Hat : 80 jobs en échec, jusqu'à 50 000 lignes de logs par job, réduits en quelques minutes à quelques causes racines. L'architecture est volontairement contrainte : **seule l'étape « Analyze » mobilise un LLM**, les quatre autres (nettoyage déterministe des logs, assemblage du rapport, création de tickets, publication wiki) restent du code Python. Bénéfices : coût d'inférence borné, sorties reproductibles, non-déterminisme confiné à la seule étape de raisonnement causal.

Sources : [Red Hat Developer Blog — How we cut CI failure triage from hours to minutes using AI](#) (29 septembre 2026) ; dépôt [Pipeline Failure Analyzer](#).



3. Argo CD 4.0 — la communauté Argo dessine l'avenir du GitOps À l'occasion d'ArgoCon North America 2026, la communauté Argo (Argo CD, Workflows, Rollouts, Events) a ouvert le **processus de vision pour Argo CD 4.0**. Les retours terrain mis en avant portent sur le passage à l'échelle (Argo CD au-delà de 60 000 applications), les topologies distribuées (« on satellites ») et l'usage d'**Argo Workflows pour le machine learning** sur de gros volumes de données.

Source : [CNCF — ArgoCon North America 2026: What to expect as the Argo community looks toward CD 4.0](#) (30 septembre 2026).

4. « Guardrails, not gates » — la policy comme primitive de plateforme (Kyverno) Article de fond CNCF : la policy doit *accompagner* (guardrail) plutôt que *bloquer* (gate). Les quatre métiers de la policy (validate, mutate, generate, verify) sont rééquilibrés : la plupart des parcs Kubernetes sont « 90 % validation » alors que **mutation et generation** portent l'essentiel du levier d'expérience développeur. Recommandations opérationnelles : messages de refus actionnables, déploiement progressif des politiques (audit → warn → enforce) et policy traitée comme un **produit** versionné plutôt qu'une configuration figée.

Source : [CNCF — Guardrails, not gates: rethinking policy in platform teams](#) (1er octobre 2026).

5. OpenClaw Enterprise — un plan de contrôle ouvert pour les agents d'entreprise Red Hat, NVIDIA, OpenAI et la communauté contribuent en open source à **OpenClaw Enterprise**, un plan de contrôle d'entreprise pour déployer et opérer des **agents persistants** multi-utilisateurs et multi-équipes (sandboxing d'agents, AI gateways, AgentOps/traçage, évaluation). En complément, une méthode de **testabilité des « skills »** d'agents est formalisée : auteur assisté (skill-creator), évaluation du déclenchement (promptfoo, assertion **skill-used**) et exécution en intégration continue.

Sources : [Red Hat Blog — Why Red Hat is building an open foundation for enterprise agents with OpenClaw Enterprise](#) (28 septembre 2026) ; [Red Hat Developer Blog — Master your skills: Building skills you can trust](#) (1er octobre 2026).

3. Impact sur les Cursus Aperta Scientia

Cursus AS200 — DevOps Platform Engineer

- **CI/CD & automatisation (D0288)** : conception de pipelines robustes et introduction d'un usage LLM *encadré* (garde-fous déterministes, coût, non-déterminisme) dans le diagnostic d'échecs de build.
- **GitOps & delivery déclaratif (D0288 / D0380)** : Argo CD à grande échelle, Argo Rollouts pour la progressive delivery et Argo Workflows pour les pipelines data/ML.
- **Gouvernance de plateforme & admission (D0280 / D0380)** : Kyverno (validate/mutate/generate/verify), conception de guardrails et déploiement progressif des politiques.
- **Exploitation conteneurs & sécurité (D0430)** : conteneurs Podman rootless, exposition maîtrisée des périphériques GPU et durcissement des runtimes OCI.

Cursus AS300 — AI Platform Engineer

- **Serving LLM & inférence distribuée (AI267)** : hétérogénéité matérielle (NVIDIA et AMD ROCm), images d'inférence supportées, benchmark GuideLLM et densité GPU.
- **Agents IA & gouvernance (AI267)** : runtimes d'agents, plans de contrôle ouverts, évaluation du comportement et tests de « skills » exécutés en intégration continue.



- **MLOps & orchestration** : pipelines versionnés (Argo Workflows) et industrialisation du déploiement de modèles sur socle Kubernetes/OpenShift.

4. Sources

- **Red Hat Emerging Technologies — Benchmarking enterprise LLMs on AMD GPUs with Red Hat AI Inference and Podman** — [Lien](#) (28 septembre 2026).
- **Red Hat Developer Blog — How we cut CI failure triage from hours to minutes using AI** — [Lien](#) (29 septembre 2026).
- **Red Hat Developer Blog — Master your skills: Building skills you can trust** — [Lien](#) (1er octobre 2026).
- **Red Hat Blog — Why Red Hat is building an open foundation for enterprise agents with OpenClaw Enterprise** — [Lien](#) (28 septembre 2026).
- **CNCF — ArgoCon North America 2026: What to expect as the Argo community looks toward CD 4.0** — [Lien](#) (30 septembre 2026).
- **CNCF — Guardrails, not gates: rethinking policy in platform teams** — [Lien](#) (1er octobre 2026).
- **Red Hat Blog — Red Hat delivers peak performance on Kubernetes and CPUs in MLPerf Inference v6.1** — [Lien](#) (22 septembre 2026, contexte).
- **Kubernetes — v1.37 release notes** — [Lien](#) ; **vLLM v0.30.0** — [Lien](#) (contexte).

Contacts & Communauté :

- **Direction des Formations** : learning@apertascientia.io
- **Référent Handicap & Bien-être (Care)** : care@apertascientia.io
- **X (Twitter)** : [@ApertaScientia](https://twitter.com/ApertaScientia)
- **Communauté Discord** : discord.gg/eDJzR5Vk85