



Veille Technologique Quotidienne

Technique

28 août 2026 – Fréquence : Quotidienne (du Lundi au Vendredi)

Synthèse & Sujet du Jour

Tech Radar & Veille : OpenShift Virt, vLLM Speculative Decoding & InstructLab

Synthèse de la veille technologique Aperta : optimisations de la Live Migration sous KubeVirt, décodage spéculatif vLLM et alignement souverain InstructLab.

1. Périmètre

Famille de veille : 2. Métiers, Compétences & Technologies.

Fréquence d'exécution : Quotidienne (du Lundi au Vendredi à 08h30).

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

2. Sujet du Jour — Article Publié sur le Radar Web

Tech Radar & Veille : OpenShift Virt, vLLM Speculative Decoding & InstructLab

Retrouvez les faits marquants de la veille technologique menée par le laboratoire d'intelligence d'Aperta Scientia pour les cursus **AS200 (DevOps Platform Engineer)** et **AS300 (AI Platform Engineer)**.

1. OpenShift Virtualization (KubeVirt) — Live Migration & Dynamic HugePages

Amélioration majeure du moteur de migration à chaud (*Live Migration*) sous OpenShift Virtualization.

- **Points clés :** Réduction de 35% du temps de convergence mémoire et support affiné des HugePages dynamiques pour les workloads hybrides (VMs d'entreprise cohabitant avec des microservices conteneurisés).
- **Impact Cursus AS200 :** Renforce le module *Virtualisation Cloud-Native (DO316 / EX316)* et l'ingénierie de résilience infra baremetal.

2. vLLM Engine — Speculative Decoding Multi-GPU & Chunked Prefill

Stabilisation de l'architecture de décodage spéculatif distribué et du *chunked prefill* sous vLLM.

- **Gains constatés :** Réduction drastique de la latence au premier jeton (TTFT) et hausse sensible du débit sous forte concurrence sur architectures NVIDIA Hopper et Blackwell.
- **Impact Cursus AS300 :** Intégration dans le module *Serving & Inférence Distribuée (KServe + vLLM v2 data*



plane).

3. InstructLab & RHEL AI — Alignement Souverain & Données Synthétiques

Nouvelle release de la suite d'outils open source InstructLab avec génération de données synthétiques multi-modèles et validation automatisée des branches de taxonomie (méthode LAB).

- **Focus :** Entraînement et fine-tuning de modèles d'entreprise sur infrastructure baremetal sur site, sans dépendance aux APIs SaaS tierces.
- **Impact Cursus AS300 :** Atelier pratique du module *Entraînement Souverain & Fine-Tuning Entreprise*.

4. OpenShift GitOps (ArgoCD) — Progressive Delivery & Custom Rollouts

Intégration native des rollouts progressifs (*canary* / *blue-green*) avec OpenShift Service Mesh et métriques SLO Prometheus automatiques pour déclencher des rollbacks déclaratifs sans intervention humaine.

- **Impact Cursus AS200 :** Standardisation des *ApplicationSets* multi-environnements avec gating automatique basé sur Prometheus Alertmanager.

5. Red Hat Advanced Cluster Security (RHACS) — eBPF Runtime Detection

Amélioration du moteur de détection d'anomalies à l'exécution basé sur eBPF dans StackRox/RHACS, ciblant la traçabilité des accès noyau et la prévention des exécutions non autorisées dans les namespaces GPU et inférence.

- **Impact Cursus AS200 / AS300 :** Modules de sécurité multi-cluster et durcissement des plateformes d'IA en production.

3. Impact sur les Cursus Aperta Scientia

Cursus AS200 — DevOps Platform Engineer

Cette veille alimente directement les modules d'ingénierie de plateforme et de sécurisation avancée de la chaîne de livraison logicielle :

- **Construction & CI/CD (D0288) :** Conception de pipelines déclaratifs Tekton, gestion des builds conteneurisés en environnement sans privilèges (*rootless*) et automatisation via Pipelines-as-Code.
- **Exploitation & Durcissement (D0380) :** Gestion multi-cluster, observabilité de la chaîne de livraison et politiques de conformité à l'échelle.
- **Sécurisation Conteneurs & Plateforme (D0430) :** Implémentation du contrôle d'admission (Sigstore Policy Controller, RHACS), signature keyless d'artefacts (RHTAS), génération de SBOM et analyse continue d'exploitabilité (TPA / VEX).
- **Certifications préparées :** Préparation opérationnelle aux examens **EX188** (Containers), **EX280** (OpenShift Administration) et **EX288** (OpenShift Development).



Cursus AS300 — AI Platform Engineer

La démarche de sécurisation de la supply chain logicielle est transposée à l'infrastructure d'inférence et aux pipelines MLOps :

- **Traçabilité des dépendances GPU & ML** : Audit de provenance et validation SLSA des images de base contenant les runtimes CUDA, ROCm et PyTorch.
- **Signature & Contrôle des Modèles** : Signature cryptographique des poids de modèles et des runtimes de serving vLLM / KServe avant déploiement en cluster.
- **Gouvernance & SBOM IA** : Inventaire automatisé des licences et détection proactive des vulnérabilités sur la pile MLOps (KubeFlow Pipelines, Ray, Triton).

4. Sources

- Red Hat Developer Blog — [Lien](#)
- OpenShift Documentation — [Lien](#)
- Kubernetes Community Documentation — [Lien](#)
- CNCF Cloud Native Landscape — [Lien](#)
- vLLM Project Documentation — [Lien](#)
- Sigstore Documentation — [Lien](#)
- SLSA Framework — [Lien](#)

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