

CHECKLIST DE PRÉPARATION À UN PIC DE CHARGE (AWS & KUBERNETES)

Cette grille répertorie les 25 contrôles techniques clés à vérifier avant une ouverture commerciale, une billetterie ou une forte variation d'affluence. Chaque contrôle est évalué selon les contraintes réelles, les objectifs de service (SLO) et l'architecture spécifique de la plateforme.

☐ À vérifier (Non encore testé ou mesuré) ☐ Conforme (Mesure validée par preuve) ☐ Écart (Risque identifié / action requise) ☐ N/A (Non applicable à l'architecture)					
RÉF.	CONTRÔLE TECHNIQUE CONCRET	PREUVE ATTENDUE & MESURE	RESPONSABLE	STATUT	CONTEXTE, LIMITES & CRITÈRES N/A
1. RÉSEAU, EDGE & INGRESS (PÉRIPHÉRIE)					
C01	Cache statique sélectif au CDN	En-tête `Cache-Control` immutable vérifié sur assets hashés ; taux de cache hit mesuré.	Infra / SRE	☐ À vérifier ☐ Conforme ☐ Écart / N/A	N/A si l'application ne sert aucun actif statique directement depuis ce domaine.
C02	Limitation de débit & WAF	Règles actives sur endpoints d'authentification et d'API ; test de tir renvoyant HTTP 429.	Sécurité / Infra	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Seuils à calibrer pour éviter les faux positifs sur les proxies d'entreprises clientes.
C03	Résilience Ingress Controller	Multi-répliques réparties sur les AZ utilisées avec PodDisruptionBudget et CPU garanti.	Infra / SRE	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Nombre de répliques dimensionné selon le volume prévisionnel de connexions concurrentes.
C04	Terminaison TLS & Keep-Alive	Déchargement TLS sur l'équilibreur ou l'Ingress ; timeouts Keep-Alive alignés avec les backends.	Infra / SRE	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Évite les handshakes répétés tout en prévenant les désynchronisations de socket HTTP.
C05	Stratégie d'invalidation ciblée	Procédure de purge par tag ou chemin documentée ; absence de purge globale aveugle.	Squad / SRE	☐ À vérifier ☐ Conforme ☐ Écart / N/A	N/A si le cache expire uniquement par TTL sans mécanisme d'invalidation événementielle.
2. COMPUTE, KUBERNETES (EKS) & AUTOSCALING					
C06	Warm pool / Pré-provisioning	Capacité de nœuds minimale rehaussée avant l'événement selon le délai de warmup mesuré.	Infra / SRE	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Optionnel si le délai d'approvisionnement mesuré sur la plateforme concernée est toléré par le trafic.
C07	Choix des métriques d'autoscaling	Arbitrage documenté entre HPA standard (CPU/RAM) et KEDA (taux RPS Ingress, lag SQS).	Squad / Lead Tech	☐ À vérifier ☐ Conforme ☐ Écart / N/A	KEDA pertinent sur flux en rafale ; HPA CPU suffisant sur charges continues prévisibles.

RÉF.	CONTRÔLE TECHNIQUE CONCRET	PREUVE ATTENDUE & MESURE	RESPONSABLE	STATUT	CONTEXTE, LIMITES & CRITÈRES N/A
C08	Requests/Limits & Répartition AZ	Ressources `requests` déclarées sur tous les pods critiques ; contraintes de topologie par zone.	Squad / Infra	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Évite le sur-engagement mémoire tout en limitant les coûts inter-AZ si le cluster est mono-AZ.
C09	PodDisruptionBudgets (PDB)	PDB configuré avec `minAvailable` ou `maxUnavailable` cohérent avec le nombre de répliques.	Squad / SRE	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Protège contre les drains volontaires ; ne protège pas contre la panne physique d'un nœud.
C10	Délai d'amorçage des conteneurs	Images conteneurs allégées ou pré-tirées ; `startupProbe` calibrée pour éviter les restarts prématurés.	Squad Backend	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Essentiel pour les applications avec temps de chargement JVM/Node long au démarrage.
3. PERSISTANCE, DONNÉES & DÉPENDANCES					
C11	Contrôle du pool de connexions DB	Plafond de connexions applicatives borné ou proxy (RDS Proxy / PgBouncer) dimensionné.	DBA / Backend	☐ À vérifier ☐ Conforme ☐ Écart / N/A	N/A si base sans serveur gérant nativement la concurrence (DynamoDB, Firestore).
C12	Dimensionnement I/O de stockage	IOPS et débit gp3/io2 mesurés lors de tirs de charge ; seuil d'alerte `DiskQueueDepth`.	Infra / SRE	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Évite la dégradation de latence lorsque la base sature sa bande passante EBS.
C13	Dérivation du flux de lecture	Requêtes lourdes de consultation isolées sur Read Replica avec tolérance au replication lag.	Squad Backend	☐ À vérifier ☐ Conforme ☐ Écart / N/A	N/A si le volume de requêtes est majoritairement transactionnel ou si le primaire suffit.
C14	Timeouts stricts sur dépendances	Timeouts explicites (2 à 4s) avec retry temporisé (jitter) ; endpoints cibles idempotents.	Squad Backend	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Prévient les blocages en cascade de threads applicatifs lors de lenteurs partenaires.
C15	Mode de dégradation gracieuse	Désactivation pilotable par configuration des fonctionnalités non critiques (recommandations).	Squad / Produit	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Permet de prioriser le tunnel transactionnel principal lors de pics extrêmes.
4. OBSERVABILITÉ & ALERTING UTILE					
C16	Dashboard unifié de command center	Écran unique regroupant RPS globaux, erreurs 5xx, latence p95, statut Ingress et santé DB.	Lead Tech / SRE	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Évite la navigation désordonnée sur des dizaines de métriques brutes en incident.
C17	Priorité aux alertes d'impact SLO	Notifications d'astreinte filtrées sur l'impact utilisateur (budget d'erreur consommé).	SRE / Lead Tech	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Réduit la fatigue d'alerte lors des fluctuations normales de charge.

RÉF.	CONTRÔLE TECHNIQUE CONCRET	PREUVE ATTENDUE & MESURE	RESPONSABLE	STATUT	CONTEXTE, LIMITES & CRITÈRES N/A
C18	Surveillance OOMKilled vs Eviction	Corrélation de l'exit code 137 avec le motif OOMKilled et les événements du conteneur/nœud, distincts des évictions.	Infra / SRE	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Permet de distinguer fuite mémoire applicative et sous-dimensionnement global de nœud.
C19	Traçage APM sur flux critiques	Taux d'échantillonnage de traces ajusté sur le tunnel d'achat sans impacter le débit réseau.	Squad Backend	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Facilite l'identification des requêtes SQL ou microservices sources d'engorgement.
C20	Quotas & Indexation des logs	Filtres d'exclusion sur logs verbeux de healthcheck ; validation des plafonds journaliers.	Infra / SRE	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Évite le bridage ou la perte de journaux de diagnostic au moment le plus critique.
5. ORGANISATION, PÉRIODE DE GEL & ASTREINTE					
C21	Période de gel des modifications	Fenêtre de freeze convenue et communiquée ; verrouillage des pipelines CI/CD de production.	Engineering Lead	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Durée adaptée aux contraintes métier (ex. 24h à 72h selon le profil de risque).
C22	Runbooks d'urgence vérifiés	Procédures pas-à-pas documentées et accessibles hors-ligne pour les scénarios de secours.	Lead Tech / Squad	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Couvre redémarrages, bascule de lecture, purge sélective et mise en maintenance.
C23	Rôle d'Incident Commander désigné	Référent identifié pour arbitrer les décisions d'exploitation ; canal de crise préparé.	Direction Tech	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Isole les ingénieurs en intervention de la pression de communication externe.
C24	Procédure de replay & DLQ	Mécanisme de rejeu depuis Dead Letter Queue testé avec traitement idempotent garanti.	Squad Backend	☐ À vérifier ☐ Conforme ☐ Écart / N/A	N/A si la plateforme n'utilise pas de files de messages asynchrones.
C25	Tir de charge de qualification	Scénarios représentatifs exécutés en environnement miroir avec marge de remédiation calendaire.	Squad / QA / SRE	☐ À vérifier ☐ Conforme ☐ Écart / N/A	Volume calibré sur les projections métier ; délai suffisant pour corriger les régressions.

Besoin d'un accompagnement d'ingénierie ? Cadrons un audit d'architecture préalable de 2 à 4 semaines indicatives.

contact@jobyops.cloud · <https://jobyops.cloud>