CRIANN

# Comité Technique HPC

CRIANN - 8 octobre 2024

# Comité technique HPC

## Ordre du jour

- Bilan d'exploitation
- Focus technique : suite de la maintenance
- Actualités MesoNET
- Agenda
- Questions / Réponses

# Bilan des attributions et des consommations

## Au 04/10/2024

- Appel à projets scientifiques 2024 (et fil de l'eau)
  - 1er AAP 2024 : 58 M.h.c attribuées / 92 projets scientifiques
  - 2ème AAP 2024 : 14,6 Mh demandées, 13,6 M.h.c attribuées / 25 projets scientifiques
    - inclut les demandes « fil de l'eau » depuis juin

- Bilan de la production d'heures 2024 au 04/10
  - Industriels : 10,4 M.h.c CPU - 1700 h.GPU
  - Académiques : 66,7 M.h.c CPU - 1,9 Mh.GPU

- Calendrier du 1er AAP 2025
  - Lancement le 14 octobre 2024
  - Fin des dépôts le 13 novembre 2024

# Valorisation des heures de calcul académiques

## Pour mémoire

- Penser à mentionner l'utilisation des moyens de calcul Criann dans les publications
  - *Ce travail a bénéficié des moyens de calcul du mésocentre CRIANN (Centre Régional Informatique et d'Applications Numériques de Normandie).*
  - *Part of this work / The present work / was performed using computing resources of CRIANN (Normandy, France)*

- Contributions au financement des coûts de calcul
  - Rappel : sur la base du volontariat !!!
  - Pour information : travail toujours en cours entre les directions des établissements et celle du Criann, sur ce sujet, et sur une évolution des statuts du Criann

# Austral

# Volet exploitation

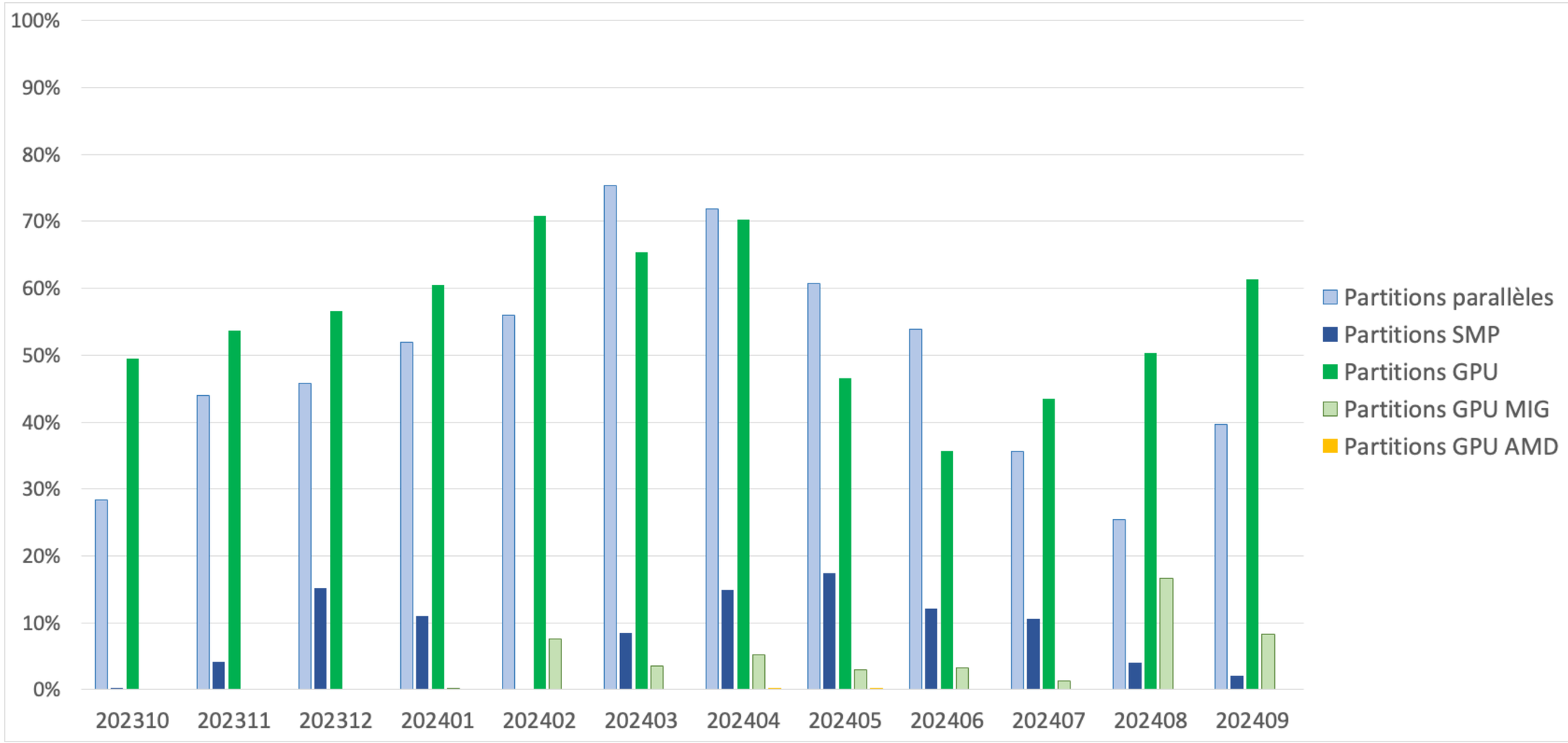
# Arrêts de production / incidents

## Depuis le dernier CT (20 juin 2024)

- Arrêt de production planifié
  - Mise à jour du 26 août au 6 sept. 2024
  - 12 jours vs 10 initialement prévus
- Arrêt pour incident
  - 22-23 sept. 2024 (1,5 jour)
  - Coupure électrique d'une durée de quelques minutes le dimanche matin. Accès aux frontales et aux données maintenu, retour en production des nœuds de calcul le lundi en début d'après-midi.

# Charge d'occupation d'Austral

Pourcentage d'occupation des ressources CPU (partitions parallèles et smp)
et GPU (partitions GPU) (hors jours de maintenance)



# Actions mises en place suite au dernier CT

## Quotas sur le stockage

- Mise en place de quotas
  - home-dir :
    - quotas pour les nouveaux comptes : 50Go
    - pour les comptes existants : utilisation + 10Go
  - dossiers de partage
    - valeur par défaut : 200Go

- Comme annoncé lors du CT de juin : les quotas sur les home-dir les plus volumineux seront revus à la baisse.

# Actions mises en place suite au dernier CT

## Partitions GPU

- Changement des durées limites pour les partitions sur GPU Nvidia
  - Partitions hpda, hpda_mig : durée max de 72h (3j)
    - Ressources max par calcul : 1 serveur (8 GPU)
    - Ressources associées aux partitions : 6 serveurs + 1 serveur MIG
  - Partition gpu, gpu_all : durée max de 48h (précédemment 72h)
    - Ressources max par calcul : 2 serveurs (16 GPU)
    - Ressources associées aux partitions : 10 serveurs
  - Ressources max par utilisateur : 16 GPU (toutes partitions confondues, hors MIG)
    - Proposition d'augmenter ce max à 24 GPU le week-end ?

# Actions mises en place suite au dernier CT

## Nouvelle partition gpu_smt

- Le SMT (Simultaneous MultiThreading) est activé sur le serveur c23gpu4
  - le nombre de cpus disponibles passe de 64 à 128
  - fait suite aux travaux sur les performances lors du chargement des jeux de données avec de nombreux petits fichiers
- Script slurm
  - similaire à ceux des autres partitions GPUs
  - partition : gpu_smt
  - demander 16 cpus par GPU
  - exemple complet disponible sur Austal : /soft/slurm/Modeles_scripts/pytorch-smt/
- Ouvert aux utilisateurs pour validation

# Sécurité

## Charte d'utilisation des moyens informatiques du Criann

- Nouvelle version de la charte et du formulaire d'ouverture de compte :
  - https://www.criann.fr/formulaires/

- CGU (conditions générales d'utilisation) :
  - https://www.criann.fr/cgu-calcul/
  - informations importantes sur les modalités spécifiques de fonctionnement du service de calcul
  - notamment politique de gestion des comptes et des données

# Focus technique :
# suites de la Mise à Jour

# Mise à jour de l’été

## Bilan

- Durée prévue de l’arrêt de service : 10j
- Durée effective pour les serveurs Fin + GPU/HPDA : 12j
- Durée effective pour les serveurs Larges + Alt : 30 j
- Raisons du retard
  - Panne matérielle sur la baie de disques
  - Problème de performances sur Openfoam 512 cœurs
    - besoin d’exécution du code en cluster dédié par l’expert HPE
- Maintenances encore en cours de finalisation…
  - Mise à jour des Bios/BMC sur les noeuds fins et les noeuds gpu/hpda
  - Arrêt successif des serveurs concernés

# Mise à jour de l'été

## Eléments mis à jour

- Système d'exploitation : RHEL 8.7 -> RHEL 8.8
- Brique de gestion du cluster : HPCM
- Réseau Slingshot : switch, firmware, drivers, fabric manager...
- Système de fichier : Lustre
- Drivers NVIDIA : 535.154.05 -> 560.35.03
  - CUDA version 12.6
- Environnement de compilation CPE : ajout de la version 24.03

# Performances post-maintenance

# Amélioration des performances du stockage

## Mis en place lors de la maintenance

- Changement du striping sur /dlocal
  - parallelisation progressive des lectures/écritures des fichiers en fonction de leur taille
- Changement du striping sur les home-dir
  - mise en place identique mais uniquement pour les futurs fichiers/dossiers
  - pour une prise en compte, recopier ou archiver/désarchiver les données
- Concrètement :
  - fichiers < 1Mo => blocs de données sur 1 seul “OST Lustre”
  - 1Mo < fichiers < 10Mo => blocs de données répartis sur 2 “OST Lustre”
  - 10Mo < fichier => blocs de données répartis sur 4 “OST Lustre”

# Validation de performance d'Austral

## Post-maintenance d'août-septembre 2024

- Le benchmark d'OpenFOAM7 de la procédure de validation a révélé des problèmes de performance, soumis à HPE

- Actions correctives
  - Correction de version de la bibliothèque libfabric du réseau Slingshot
  - Ajout de procédures de « drop_cache » et de « compact_memory », dans le pré-traitement (« prolog ») des travaux Slurm exclusifs

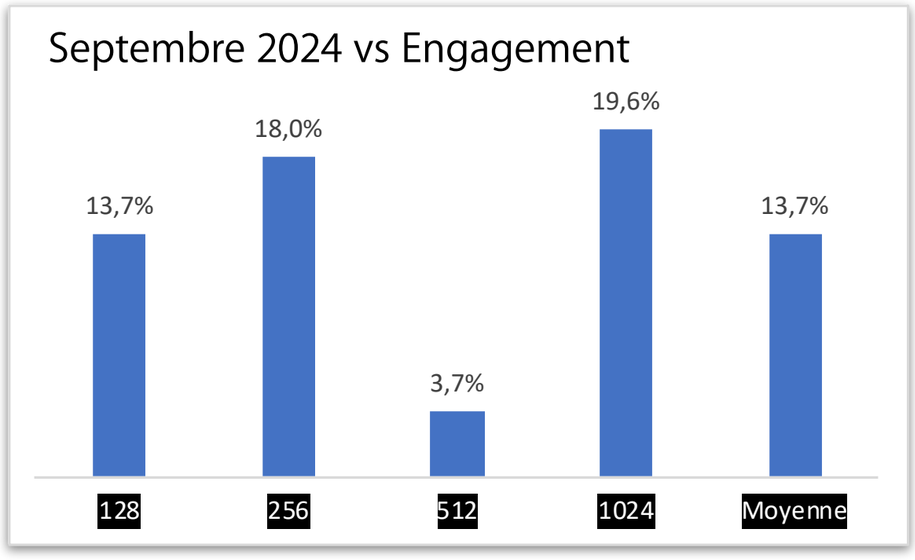
# Validation de performance d'Austral

## Post-maintenance d'août-septembre 2024, OpenFOAM7

**Cas 1024 cœurs**

Engagement constructeur : 1930 sec
Recette (août 2023) : 1676 sec
Février 2024 : 1524 sec
Septembre 2024, installation post-maintenance comprenant libfabric/1.20.1 : 3739 sec ❗
Septembre 2024, libfabric/1.23.1 : 1786 sec
Septembre 2024, libfabric/1.23.1 + drop_cache + compact_memory : 1552 sec ✅

**Cas 128 à 1024 cœurs : variation relative de WallClock (pourcentage positif = gain)**




CRIANN

# Logithèque

# Développement de la logithèque

## Installations récentes

- Visualisation (structures cristallines et moléculaires)
  - Xcrysden 1.6.2

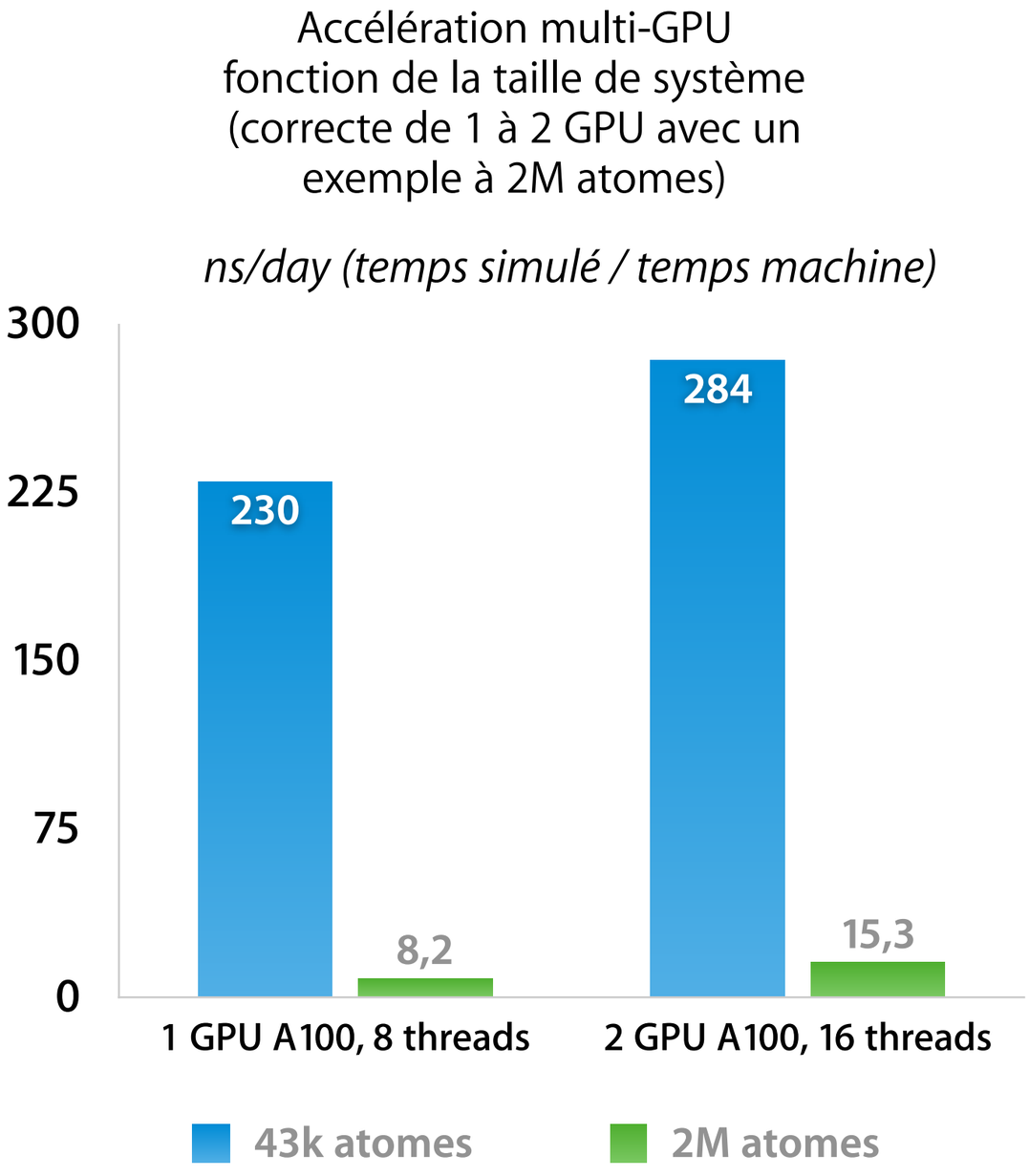
- CFD
  - OpenFOAM v2406

- Simulation atomistique
  - WIEN2K (privé)
  - Gromacs 2024.2

- IA Deep Learning
  - Tensorflow 2.16.1 (suite à mise à jour des drivers nvidia)

# Simulation atomistique

## Gromacs 2024.2




- Commandes utiles sur Austral
  - module avail atomic_simu
  - module help atomic_simu/gromacs/2024.2

- Options de configuration
  - mono-serveur, multi-thread, multi-GPU

```
GROMACS version:     2024.2
Precision:           mixed
Memory model:        64 bit
MPI library:         thread_mpi
OpenMP support:      enabled (GMX_OPENMP_MAX_THREADS = 128)
GPU support:         CUDA
NBNxM GPU setup:     super-cluster 2x2x2 / cluster 8
SIMD instructions:   AVX2_256
CPU FFT library:     fftw-3.3.8-sse2-avx-avx2-avx2_128
GPU FFT library:     cuFFT
```

# MesoNET

| Site / machine | Descriptif | Dispo. |
|---|---|---|
| Calmip Toulouse<br>*Turpan* | 15 nœuds, interconnect 2xHDR, stockage NFS<br>Total 1200 cœurs ARM 3Ghz et 30 GPU Nvidia A100 80 Go | disponible |
| Criann Rouen<br>*Boreale* | 9 nœuds vectoriels x 8 VE NEC SX Aurora Tsubasa 20B + 2 Intel 16 cœurs 2.9 GHz<br>Interconnect IB 200 Gbps – Stockage GxFS 510 To utile | disponible |
| Romeo Reims<br>*Juliet* | 3 nœuds IA<br>x 8 GPU A100 Nvidia 80 Go + 2 AMD EPYC 7663 56 cœurs 2 GHz<br>Interconnect IB 200 Gbps + 10 Gb eth | disponible |
| Strasbourg<br>*Vesta* | 3 nœuds GPU AMD<br>x 10 GPU AMD Mi210 64 Go + 2 AMD EPYC 7643 48 cœurs 2.3 GHz<br>IB 100 Gb + 25 GbE<br>Stockage 212.8 Tio nets SATA | disponible |
| Glicid Nantes<br>*Phileas* | 32 nœuds x 2 CPU Intel SPR 48 cœurs 2.1 GHz<br>Interconnect IB 100 Gb + 25 Gb eth<br>Stockage GPFS 285 To | à venir |
| Lille<br>*Zen* | 72 nœuds x 2 AMD EPYC Genoa 9534 64 cœurs 2.45 GHz<br>OmniPath 100 Gb + ethernet 10 Gb<br>Stockage BeeGFS 1 Po utile | disponible |
| Marseille | GPU H100 | à venir |
| Grenoble | Openstack | à venir |

# MesoNET

## Avancées du projet

- De nouvelles machines sont mises en production progressivement
- Documentation
  - https://www.mesonet.fr/documentation/user-documentation/
  - Boreale et Turpan : rubrique architectures spécialisées
  - Autres : machines code-formation

# MesoNET

## Demandes d’accès aux ressources

- Accès à MesoNET :
  - https://www.mesonet.fr/documentation/user-documentation/acces/portail
    - Compte à demander sur https://iam.mesonet.fr/login
    - Puis dépôt de projet sur le portail https://acces.mesonet.fr

- Cas particulier de la machine Boreale du Criann
  - e.g. 1000 heures x VE (Vector Engine) à demander pour Boreale
  - si besoin de plus amples informations, écrire à support-boreale@criann.fr

# Agenda

# Formations

## À venir

- Conférence MesoNET « **État de l'art du calcul quantique** » par Eviden
  - le 9/10 à l'Insa et le 10/10 au Greyc
- **Python pour le HPDA -** 12 et 13 décembre
- **Prise en main d'Austral pour un usage généraliste** le 15/10 (visio)

• En cours d'organisation

- **CPE** (Cray Programming Environment), par HPE
  - Outils de compilation, débugage, profilage
- **MLDE** - Machine Learning Development Environment, par HPE

# Agenda

- JCAD 2024, Bordeaux, du 4 au 6 novembre
  - Inscriptions closes
  - Événement diffusé en ligne
  - https://jcad2024.sciencesconf.org/resource/page/id/11
- Prochain Comité Technique
  - Q1.2025
- Organisation d'une journée scientifique des utilisateurs en 2025
  - Propositions de sujets, dates ?

Le plateau de calcul intensif du Criann est cofinancé par la Région Normandie, l'État français et l'Union européenne (Fonds Feder).
MesoNET bénéficie d'un financement de l'Agence nationale de la recherche au titre des Investissements d'avenir.
Le Centre de Compétence EuroCC français est cofinancé par l'Union européenne et par l'État français.
Le réseau régional Syvik est cofinancé par la Région Normandie et par l'Union européenne (fonds Feder).
Le fonctionnement du Criann bénéficie du soutien de la Région Normandie.

RÉGION NORMANDIE

RÉPUBLIQUE FRANÇAISE
*Liberté*
*Égalité*
*Fraternité*

UNION EUROPÉENNE
Fonds européen de développement régional

CRIANN

Centre Régional Informatique et d'Applications Numériques de Normandie
www.criann.fr