Designing Alice Recoque's Storage for Exascale AI

Thomas Leibovici, CEA

ALICE RECOQUE
BEYOND HPC



AI Factory France

Alice Recoque Supercomputer



- 2nd Exaflopic system in Europe (the first was JUPITER, hosted by JSC)
- 1st Exaflopic system in France
  - Peak performance > 1 ExaFlops linpack

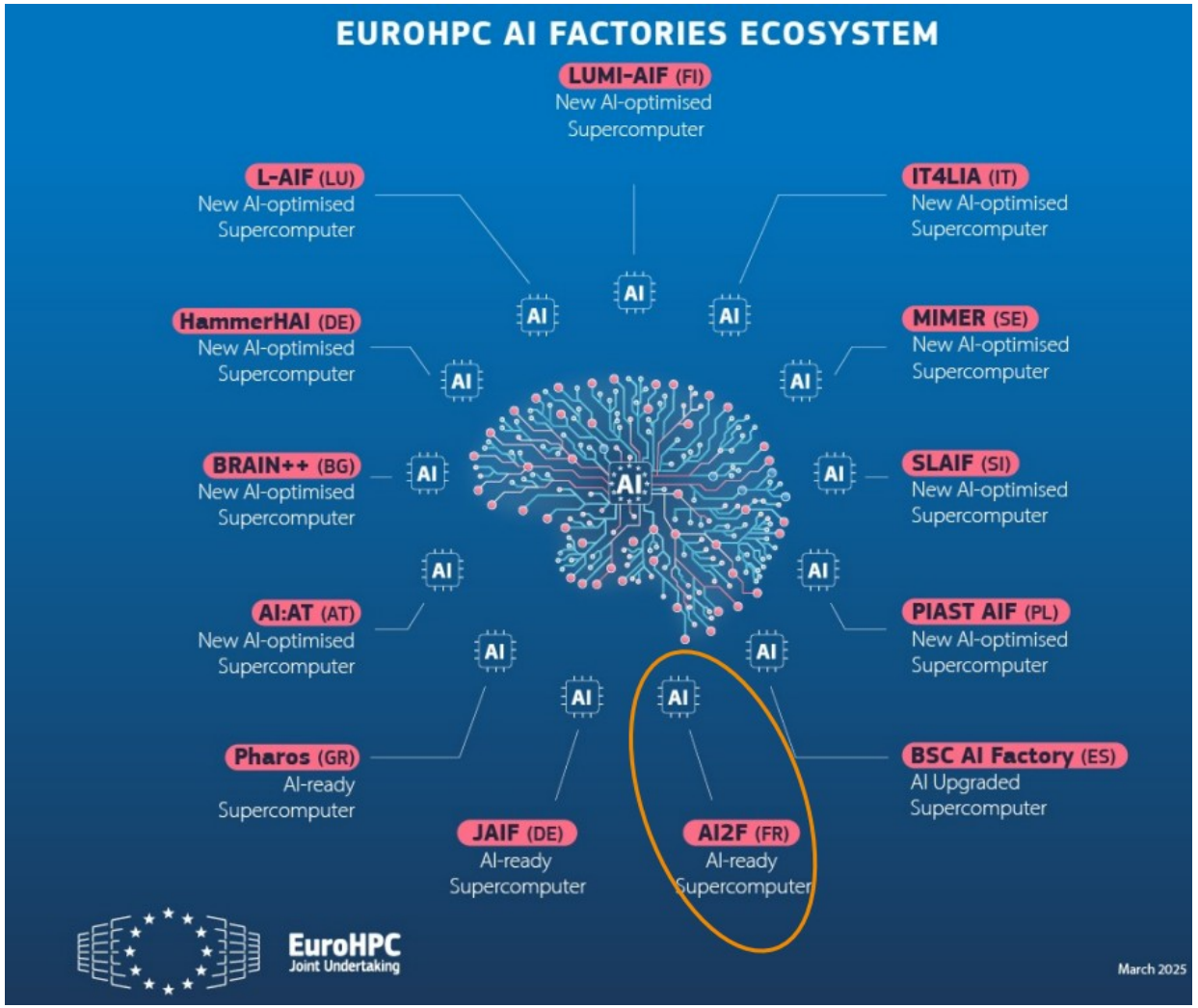
- Led by the Jules Verne consortium including France (GENCI, CEA), Netherlands (SURF) and Greece (GRNET).
- Built by Bull, based on BullSequana XH3500
- Hosted by the *Très Grand Centre de calcul du CEA* (TGCC)
- Operational launch: 2027

- Alice Recoque is a key component of AI Factory France (AI2F)

ALICE RECOQUE
BEYOND HPC



IA Factory France



MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'ESPACE
Liberté
Égalité
Fraternité

Accueil > Presse > Communiqués de presse

Publié le 11.03.2025

AI Factory France : la France lauréate du programme européen EuroHPC

Les ministres Philippe Baptiste et Clara Chappaz se félicitent de la sélection du projet AI Factory France dans le cadre du programme EuroHPC. Ce choix confirme l'excellence française en intelligence artificielle et en calcul haute performance. Cette reconnaissance européenne renforce la dynamique impulsée par le Sommet IA et positionne la France comme un acteur incontournable de l'IA dans un contexte mondial en pleine mutation.

IA Factory France

- One-stop shop for access to AI resources in France
- Provides computing resources, expertise, and training
- For researchers, startups, SMEs, large companies, public services, etc.
- Official website : https://aifactory-france.eu



Main AI Workloads: Training

read

- **Intensive read workloads on huge datasets**
  - 100s of Terabytes
  - Random access (data must be shuffled to avoid training bias)
  - Large sequential access (TFRecord, Parquet…)

write

- Checkpoints : **large outputs** (GB to TB)
- Logs, training metrics
- Final output: model weights

Requirements: high bandwidth, low latency (especially for metadata)

Main AI Workloads: Inference

read

- Loading of pre-trained models
- Intermittent read of models, if they do not fit entirely in memory
- Use of caches (Memory, HBM, NVMe) to avoid slow HDD accesses, especially for large models

write

- Logs
- Storage of results
- Cache population

Latency-critical

Alice Storage Focus for AI

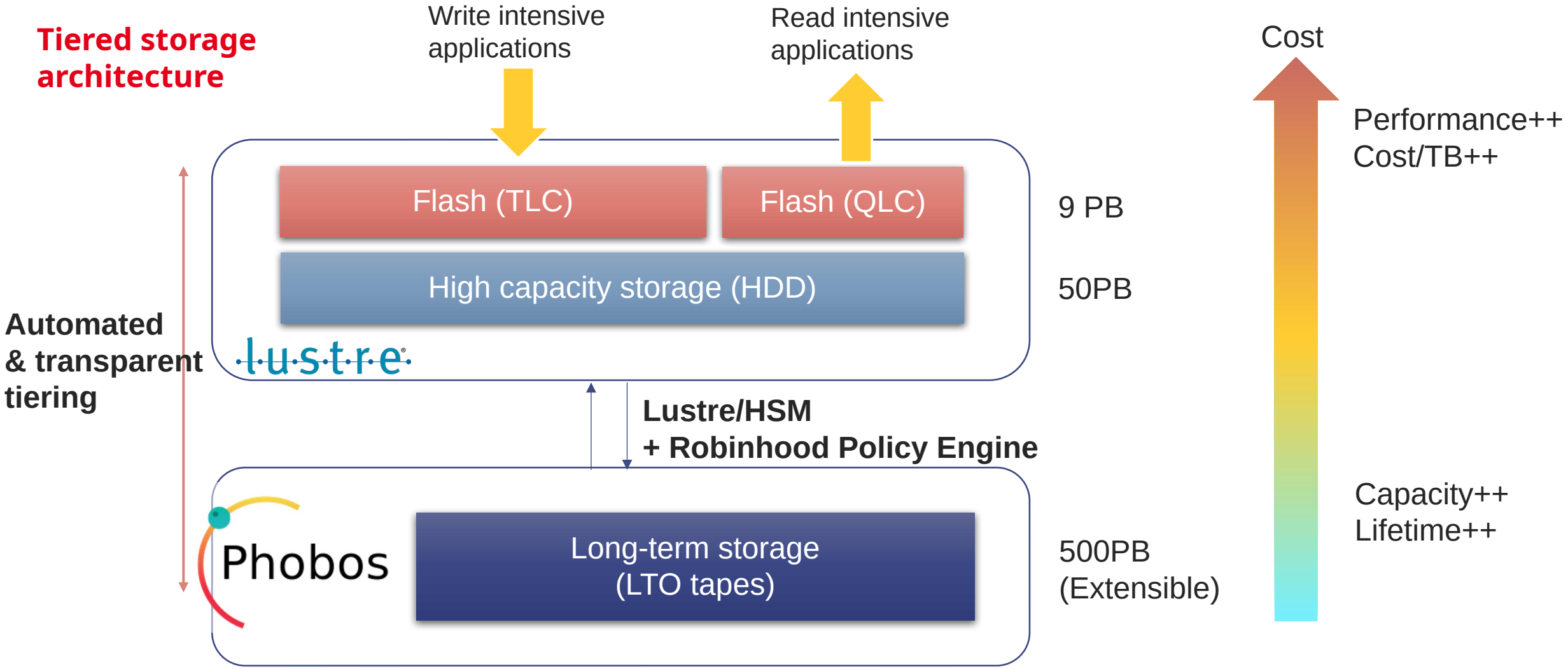
- **Outstanding metadata performance**
  - 32 metadata servers
  - Can ingest 1.5M file creation/sec
  - Can retrieve medatada (stat) of 3M files/sec

- **High throughput**
  - Commitments for TLC flash storage:
    - Write : 1.5 TB/s throughput, 6 MIOPS
    - Read : 2.25 TB/s throughput, 83 MIOPS

- **S3 service**

Taming (Exorbitant) Media Costs

Tiered storage architecture



Ephemeral storage

- Tool: **Bull ARGOS Data Management – GBF (Global Bunch of Flash)**
- Provides **On-demand Dedicated Lustre Filesystems**
- Benefits :
  - Isolate job IOs to prevent interference
  - Make I/O performance repeatable
- Integrated with SLURM



Specific Benchmarks

- **MLperf Training** (llama2 70B, RetinaNet, BERT-large, GPT3...)
- **MLperf Inference** (llama2 70B, Mistral 8x7B, stable diffusion, RetinaNet...)
- **MLPerf DLRM, stable diffusion, R-GAT…**
- **Alphafold 2**
- **MLperf storage**

Prospects

- **Large consistent filesystems** is not required for AI
- **Key-Value Stores** and **object stores** can do the job
- S3 Object Storage is commonly used, but common solutions suffer from high latency
- Emerging S3/RDMA
- **AI requirements will drive the emergence of new technologies that will be beneficial for storage and HPC in general**
  - Interconnects: Nvlink, Infinity Fabric
  - Storage for high-performance bus
  - HBM memory
  - GPU Direct I/O


TGCC