SUPERPOWERING THE EDGE

CAPE

European Open Compute Architecture for Powerful Edge

Fred Buining CTO, HIRO MicroDataCenters | CAPE technical coordinator

HiPEAC | 12 partners | December 2024 to November 2027 | Approx. €6 million EU funding

CAPE CONSORTIUM | ADMINISTRATIVE LEAD: BIELEFELD | TECHNICAL LEAD: HIRO

Universität Bielefeld | HIRO MicroDataCenters | PCB Design | paraXent
Ryax | CAD-Terv | Christmann | Fraunhofer ITWM
IPTO | EPFL | Universität St. Gallen | PICMG

Funded by the European Union, Horizon Europe Grant Agreement No. 101189899

Source: CAPE consortium | CORDIS project record

Intelligent environments emerge through interactions

DIGITAL ENVELOPES

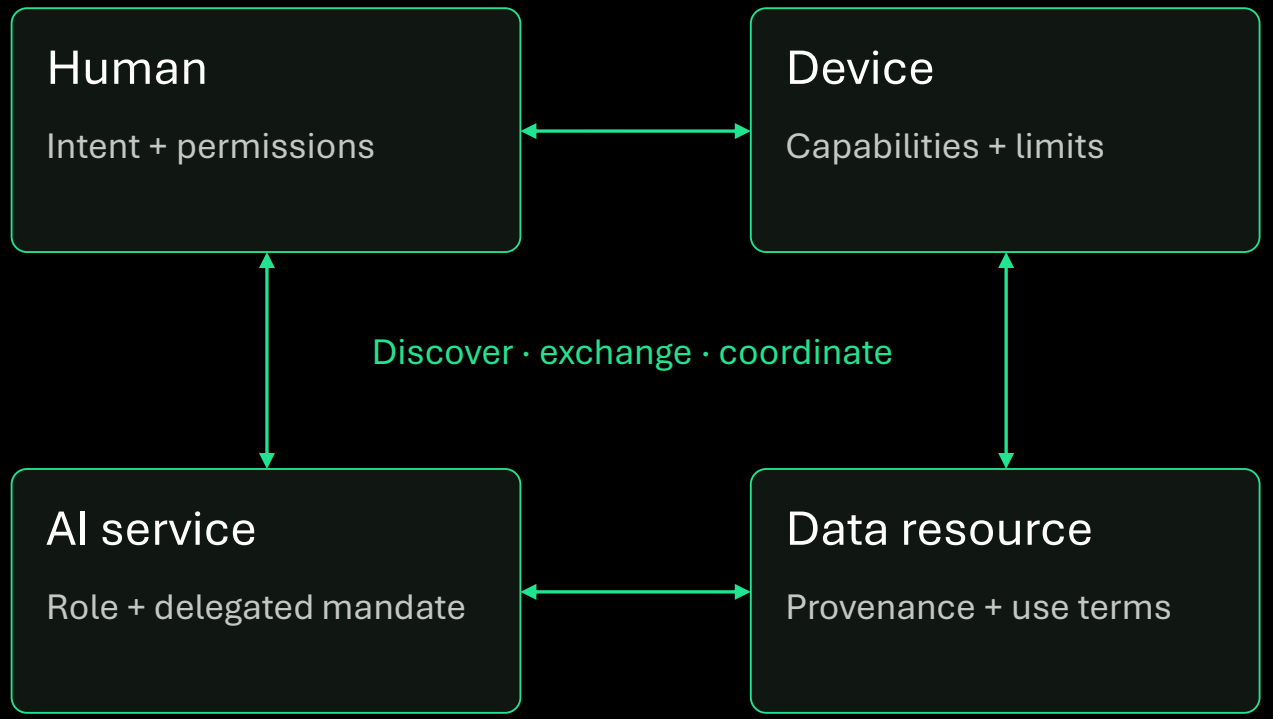
A digital representation of a person, device, AI service or data resource.

WHAT EACH PARTICIPANT BRINGS

Identity, capabilities, permissions and conditions for participation.

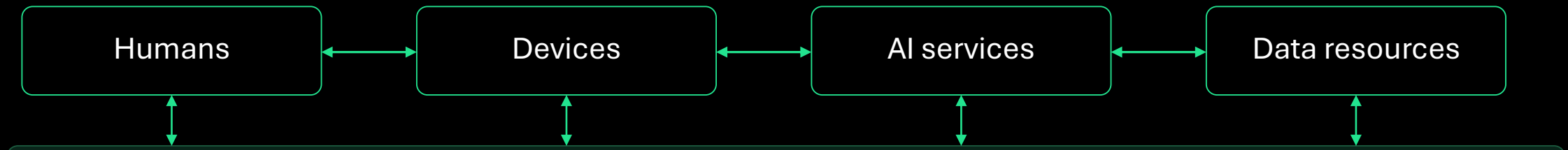
WHAT THE INTERACTION NEEDS

Relevant context, available services and resources to decide and act.



Note: Conceptual interaction diagram; no central controller.
Source: HiPEAC Vision 2026, pp. 20 to 24 | IEEE 2874-2025

Local services support interaction without taking over



LOCAL SUPPORT SERVICES

Locally operated · open · configurable

CONTEXT + AI

Provide relevant data and suitable AI. Preserve provenance, freshness and use conditions.

MANAGEMENT + GOVERNANCE

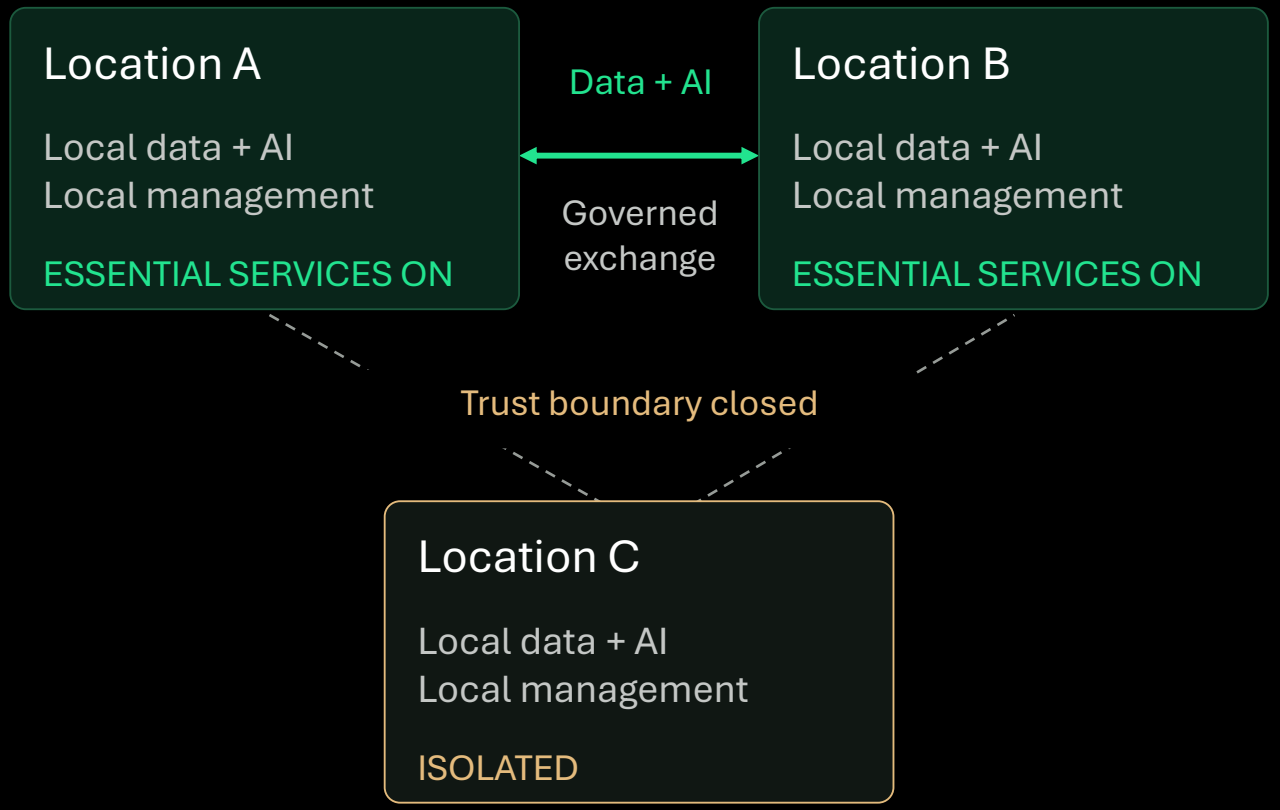
Discover capabilities and permissions. Monitor alignment. Intervene within delegated authority.

STATE CONTINUITY

Preserve relevant context, commitments and progress as participants, services or location change.

Note: Proposed support architecture; credentials alone do not provide authorisation.
Source: W3C Verifiable Credentials v2.0

Autonomous locations cooperate without sharing every failure



BOUNDED DEPENDENCIES

Keep essential capabilities local.
Contain failure and security impact.

EAST-WEST PRIORITY

Design bandwidth, traffic isolation and recovery capacity for SLA/SLOs.

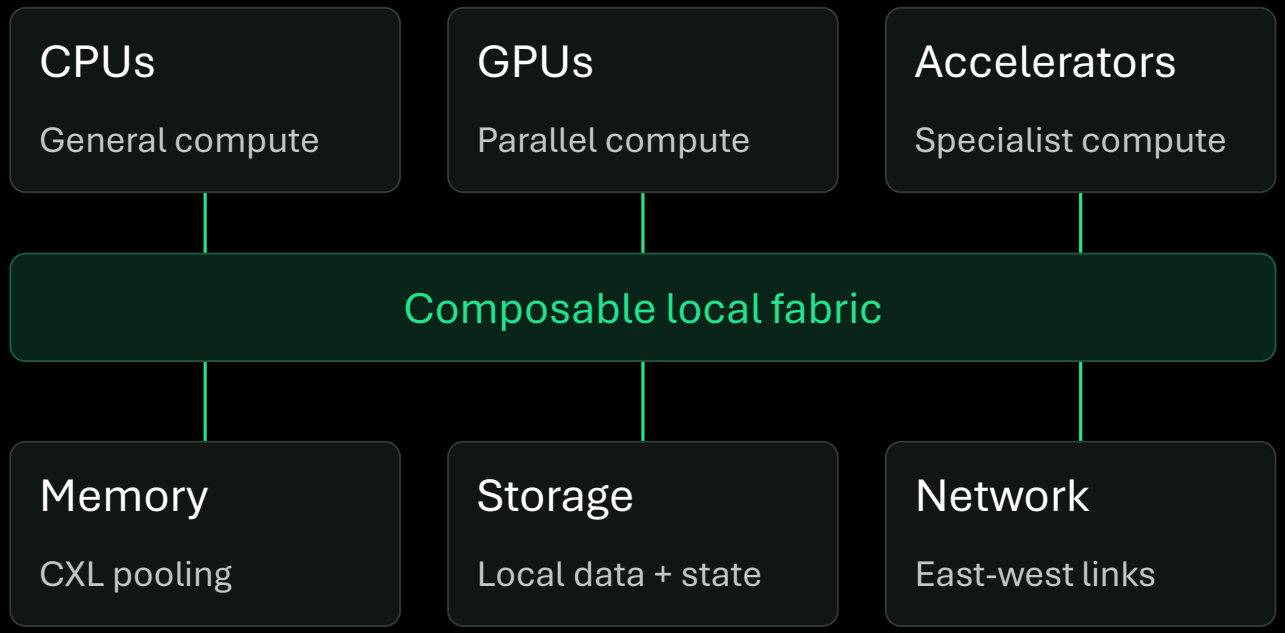
EXPLICIT CONTINUITY

Define what continues, degrades or pauses when a dependency fails.

Note: Design objective, not a guarantee that a mesh alone delivers resilience.
Source: Local-first software | Keyhive research

Composable infrastructure makes local deployment adaptable

LOCAL RESOURCE POOLS



Within a site: compose resources
Between sites: federate services and data

Note: CXL supports memory pooling; the broader resource mix is an architectural design.
Source: CXL 3.0 white paper

RIGHT-SIZED MIX

Fit local services, devices, power and cooling conditions.

USEFUL UTILISATION

Reduce idle capacity while retaining failure headroom and service guarantees.

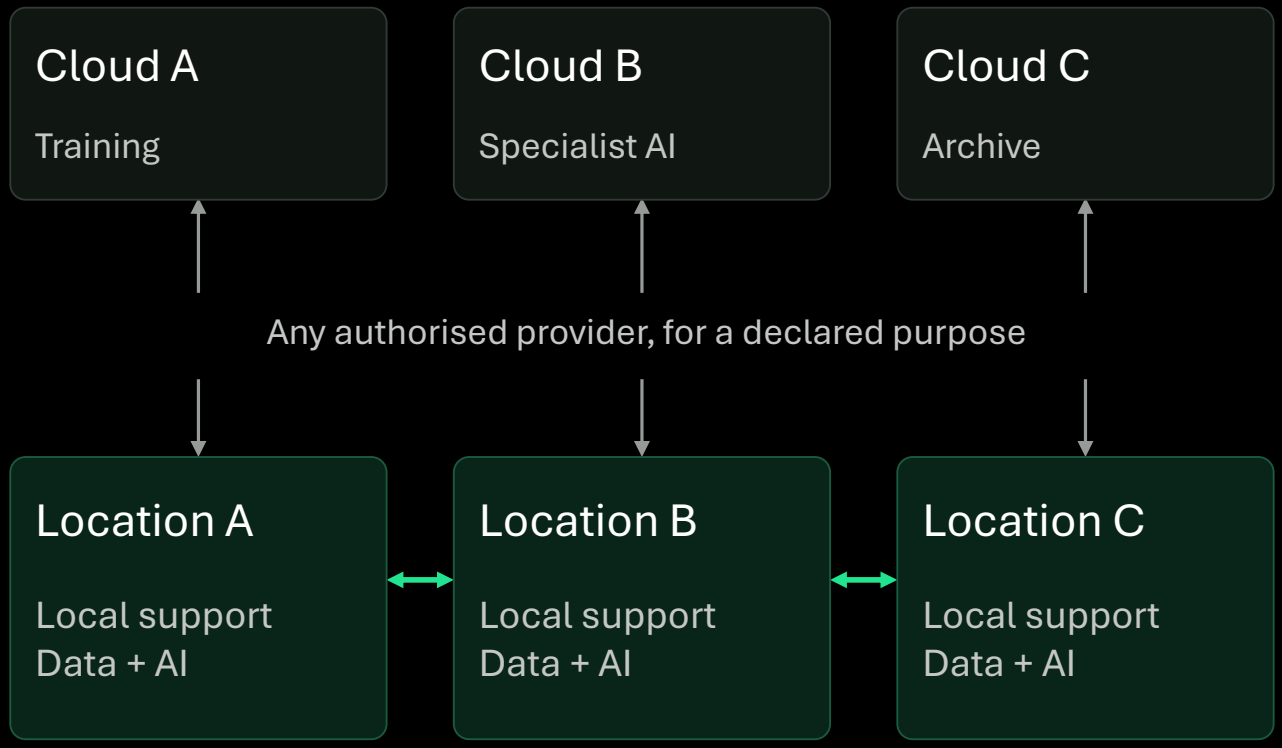
MODULAR UPGRADES

Expand or replace resource pools without replacing the whole system.

LOW-TOUCH LIFECYCLE

Secure remote updates, rollback and recovery with few site visits.

Connect to distant clouds without giving up local autonomy



Autonomous locally · cooperative with neighbours

CAPE DEVELOPS THE FOUNDATIONS

CXL-based composable hardware

Cloud-agnostic
Infrastructure from Code

Governance of data, models
and applications

2.7× Projected global datacentre capacity demand

2025 to 2030 | 82 to 219 GW

Note: CAPE foundations, not a claim that the full interaction environment is delivered.
Source: CAPE / CORDIS | McKinsey, May 2025 forecast

Powerful Edge-Cloud as a Service (PEaaS)

Enterprise digital sovereignty from three loosely coupled services



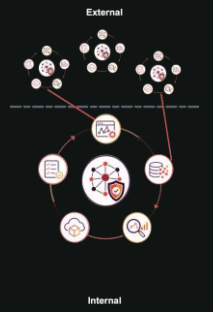
Sovereign Edge-Cloud Infrastructure as a Service

- Resilient mesh of heterogeneous EdgeMicroDataCenters
- Intelligent infrastructure as code for developers
- Orchestration across on-prem and off-prem resources
- AI-supported DevOps (POLAR) to find and fix failures



Sovereign Dataspaces as a Service

- Governed data exchange between internal sites and external partners
- Data stays under the control of its owner
- Policies decide who may use which data, where and for what



Sovereign AI-Management as a Service

- HIRO-CF: open-source platform to manage and govern AI/ML models
- Registration, training, deployment and monitoring
- Audit trail, transparency and explainability built in

GOVERN where AI, data and workloads may execute

DECIDE keep local, use enterprise infrastructure or extend to sovereign or external cloud

ENFORCE identity, sovereignty and governance policies across the continuum

Why CAPE, GenAI trends

Models, hardware and the nature of work are converging

MODELS

One training run yields a family. Smaller models are cut out of the large one, or the large model teaches small ones to reproduce its answers.

Mixture-of-experts models are sparse. A router activates only a few per token. All expert weights still need storage. They may stay on GPUs; CPU offloading is optional.

Models age in months. Each generation is better at the same size. Value moves from renting tokens to owning data, routing, evaluation and control.

HARDWARE

Inference splits into stages. Each stage stresses something else: arithmetic, memory bandwidth or memory capacity. No single GPU fits all, so hardware becomes heterogeneous: CPU, GPU, accelerators, memory pools.

Memory becomes the engine. Processing long documents, codebases, tools takes seconds, minutes. Caching the result (KV Cache) and reusing costs far less.

Many machines, one model. Split by layers: little network traffic, buys memory. Split inside layers: buys speed, needs a fast network.

WORK

Work becomes agentic. Many simple, routine steps and few hard ones. Cost and speed scale with the number of steps.

The central route hits limits. Memory, GPU and energy prices and lead times rise. Large central capacity is expensive, slow to build and hard to power.

A TYPICAL AGENT TASK, ILLUSTRATIVE



routine step: small model

hard step: large model

Smaller, specialised models, running on mixed hardware, close to the work.

Converging towards the enterprise

Right-size every step: small models for routine work, large models only for the hard steps

FRONTIER ROUTE

- Costly to train and costly to use at scale.
- Even with falling token prices, a frontier model for every step is uneconomical.
- Frontier companies build their own infrastructure, for example Anthropic's $50 billion data centre programme.
- They build vertical offerings with consultancy partners, for example in pharma and life sciences.

ROUTING

- Most agent steps are routine. A small model handles them.
- Hard steps escalate to a larger model. A router sets the thresholds.
- Model families come in smaller rungs from one training run.

40 to 70%	of calls in three open agents could move to small models
10 to 30x	cheaper per call for a small model
up to 98%	cost cut by cascades, in benchmark studies

ENTERPRISE ROUTE

- Small and mid-size open models, retrained on data close to the process steps they serve.
- Done by the enterprise itself or by a service provider.
- Possible once they have the process and the tools.

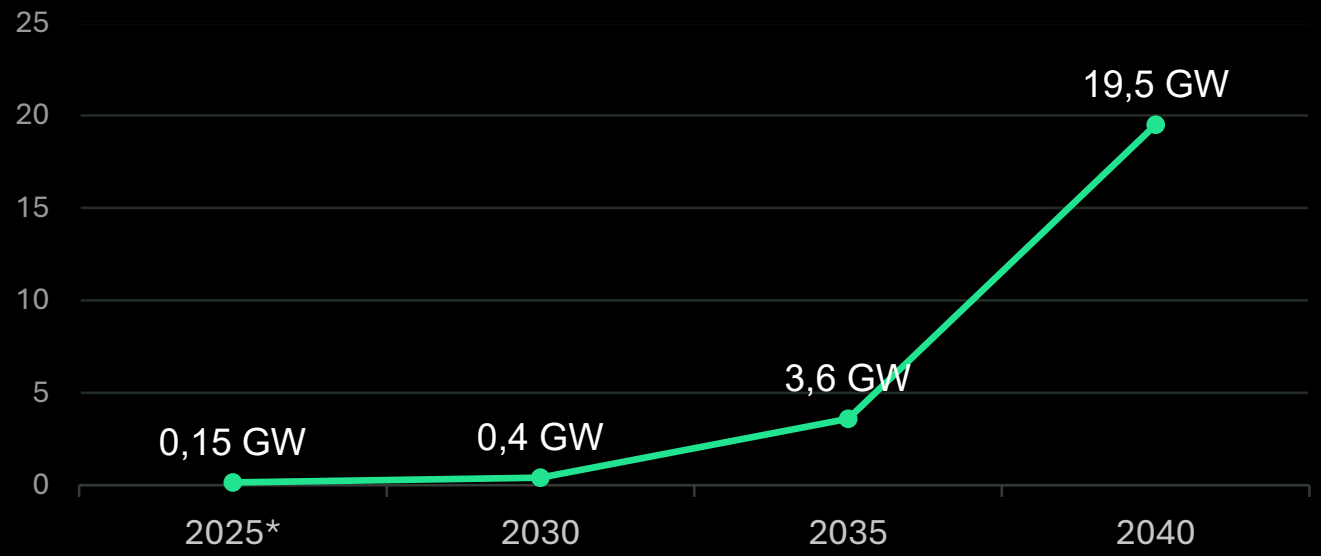
CAPE develops the open infrastructure underneath this route.

CAPE connects local control with cloud scale through open hardware and an open edge-cloud software stack.

Source: Anthropic infrastructure | Life sciences | NVIDIA SLM study | FrugalGPT | CAPE

GenAI could create a €140bn annual European local edge-cloud market

EU27 INSTALLED IT CAPACITY | HIRO SCENARIO



	2025*	2030	2035	2040
Locations	5,000	11,500	60,000	200,000
kW / site	30	35	60	≈100

2040 ANNUAL MARKET SPENDING

≈€110bn recurring

Data + AI platforms, operation and infrastructure-as-a-service

≈€25bn new systems

Customer-owned infrastructure

AI-INTENSIVE, VENDOR-NEUTRAL

60% accelerators | 40% CPU

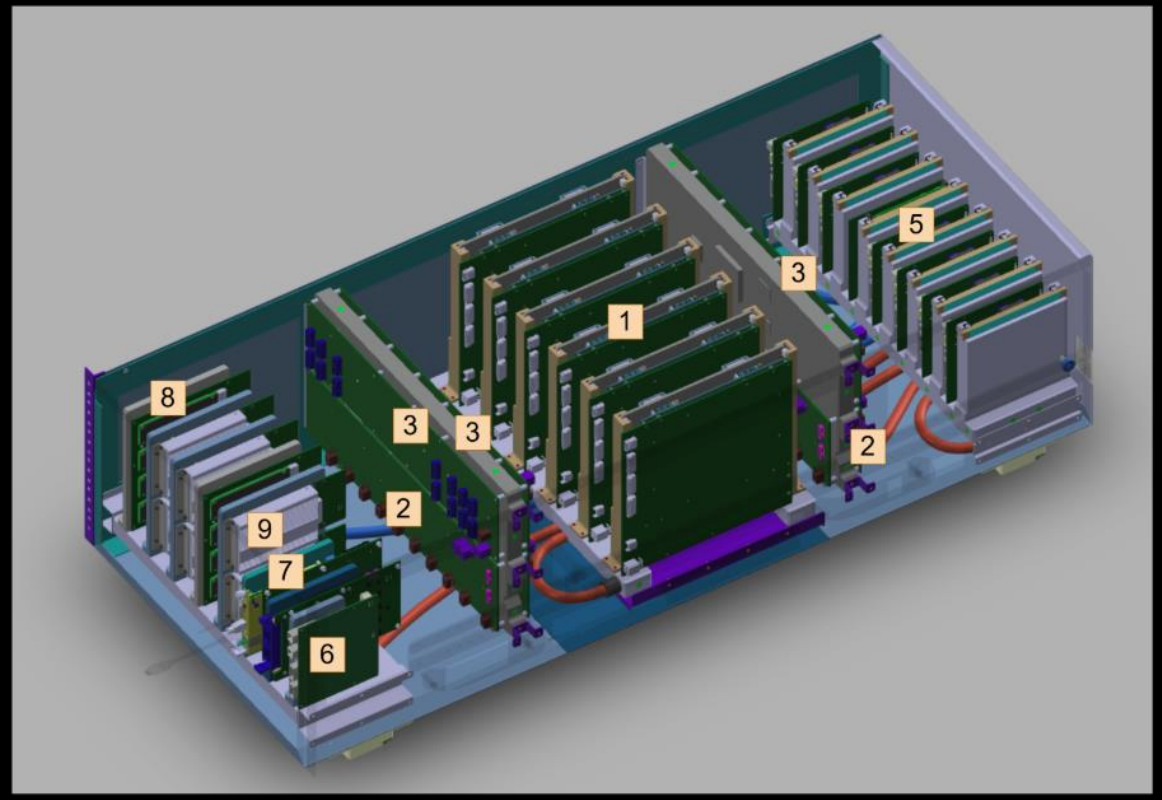
Compute power split; plus CXL memory, storage and east-west networking

Entire market, not HIRO revenue. HIRO scenario, not an external consensus forecast.

Assumes 50% fully managed and 50% real-price compression vs 2026. Constant 2026 euros; energy/connectivity excluded. *2025 assumed.

EdgeMicroDataCenters (EMDC)

Open standards for vendor-agnostic, heterogeneous edge-cloud infrastructure



(1) COM-HPC (2) Broadcom Trident Ethernet switch, 2x100 GbE uplink (3) PCIe/CXL switch (4) 48V to 12V converter (5) PIC64-HPSC or NVIDIA Thor modules (6) OCP-NIC 3.0 (7) OCP-BMC DC-SCM 2.1 (8) EDSFF E1.S SSDs (9) EDSFF E3.S CXL memory

Modular by industry standard

- Nodes on COM-Express (180 W) or COM-HPC (400 W)
- Several nodes per module, 1 to many modules per EMDC

Heterogeneous by design

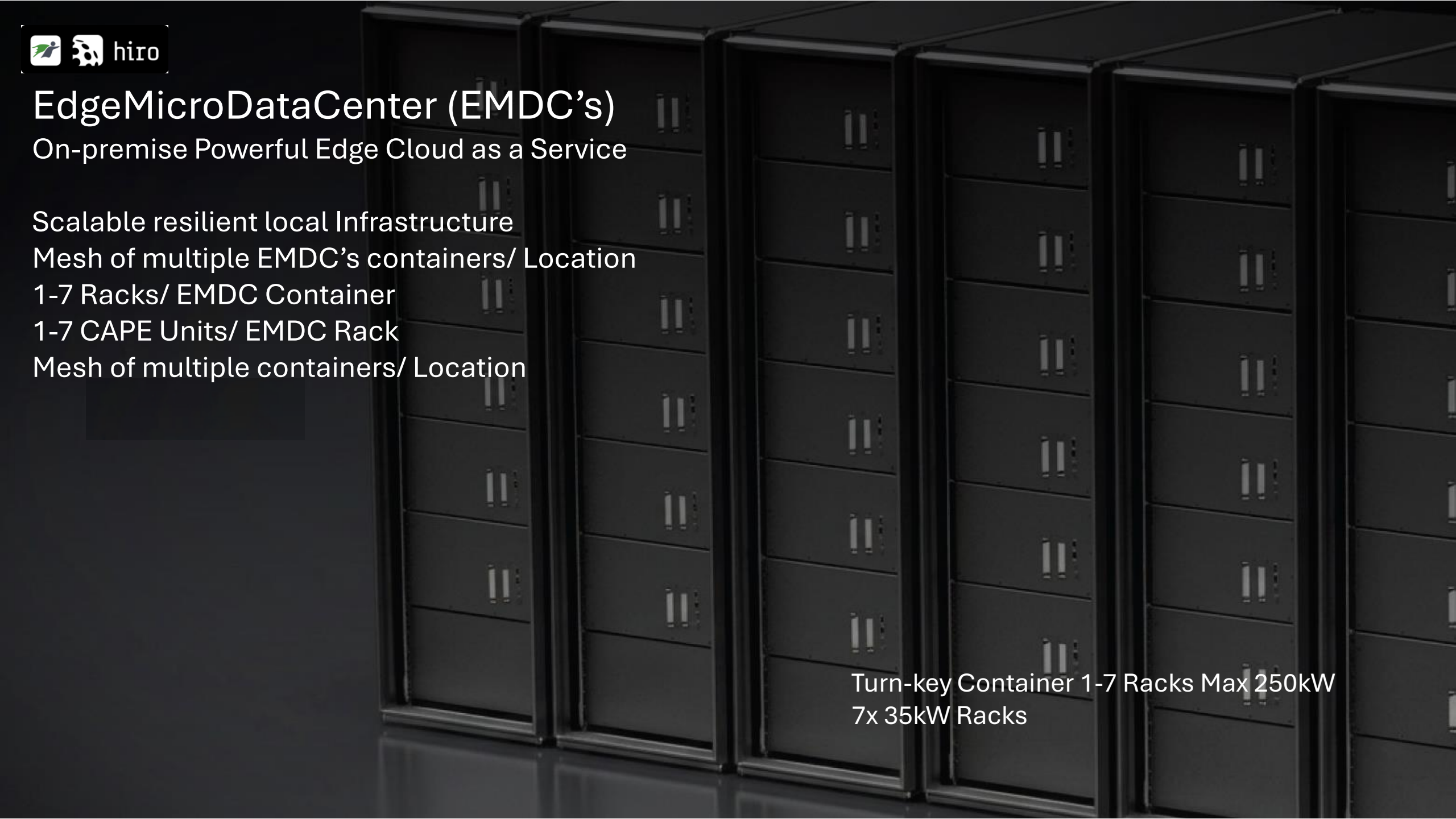
- Purpose-composed mix of CPU, GPU, FPGA, RISC-V, neuromorphic
- CXL memory for edge AI, dual fabric Ethernet and PCIe

Solid state, efficient phase change cooling

- No moving internal parts
- Internal phase change, external high grade warm water to facility or air

Autonomous and federated

- Each EMDC runs AI and data locally
- Federates with other EMDCs and connects to any distant cloud



EdgeMicroDataCenter (EMDC's)

On-premise Powerful Edge Cloud as a Service

Scalable resilient local Infrastructure
Mesh of multiple EMDC's containers/ Location
1-7 Racks/ EMDC Container
1-7 CAPE Units/ EMDC Rack
Mesh of multiple containers/ Location

Turn-key Container 1-7 Racks Max 250kW
7x 35kW Racks

Energy Efficient
Low latency Sovereignty

Industry standard form factors
European OEM’s, design and manufacturing

Energy efficient through
- Intelligent orchestration
- Compact form factors
- DC Power
- Warm water cooling

Cooling technology: Two-phase direct-to-chip
Liquid heat removal target: ≥90%
Facility-water: 40°C supply / 50°C return
Design ΔT: 10 K

Max 7x CAPE Units Rack Tot. 35kW



POWER
AC-DC or Fully DC powered
Wind, Solar, Battery, Hydrogen

COOLING
- Two phase direct to chip
- Warm facility water
  inlet 40°C outlet 50°C

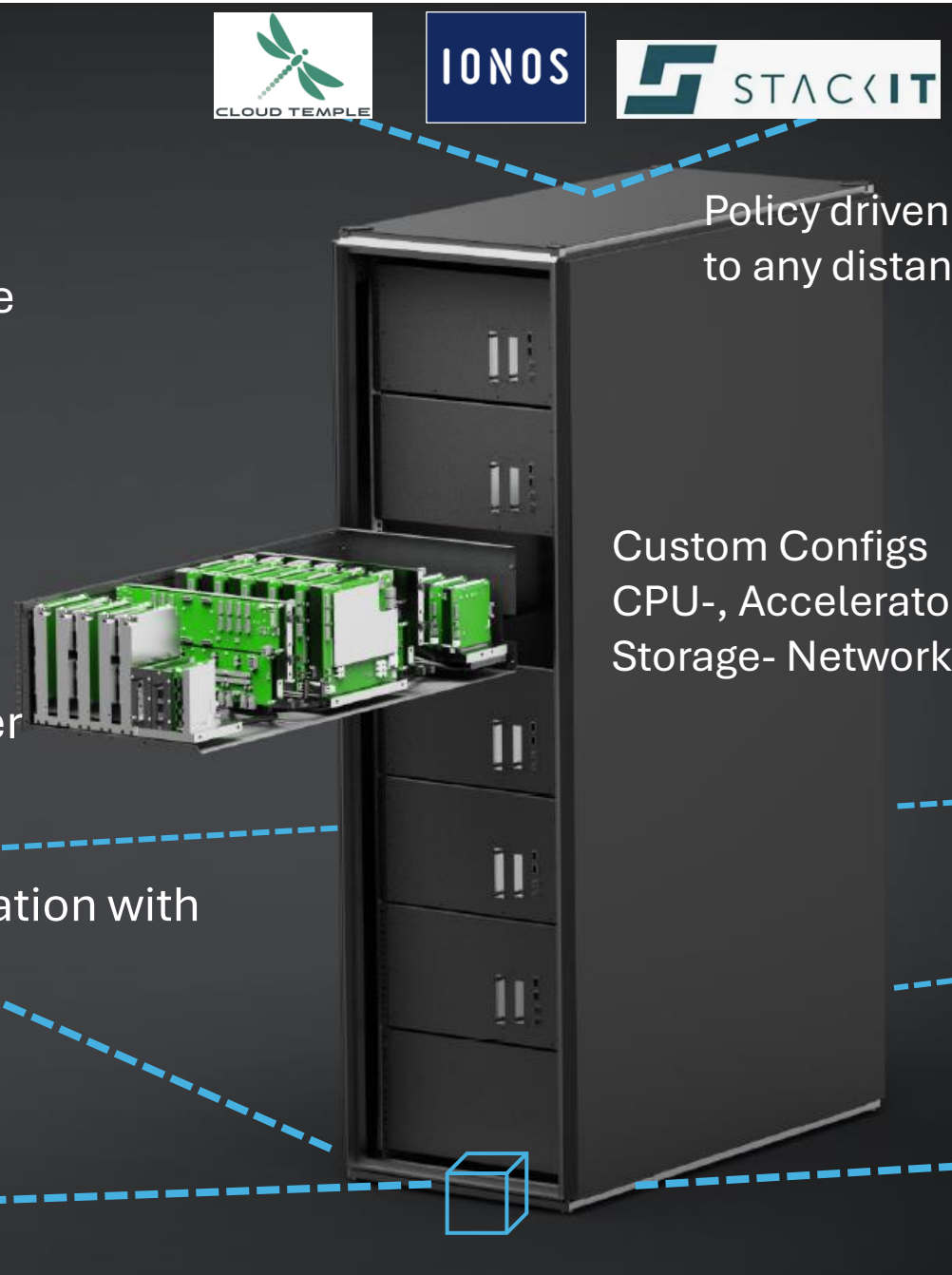
Zoom in one rack of of CAPE

Composable EMDC’s

At runtime configuration of right sized hardware
High utilisation levels of hardware

Each EMDC has 1-7 CAPE units
CAPE unit has
- Pools of CPU’s, FPGA’s, GPU’s Storage
- Dual fabric (Ethernet, PCIe) Networking
- Runtime Intelligence creates the ideal server for each workload

Each EMDC is autonomous and works in federation with
- Other EMDC’s
- Distant cloud providers

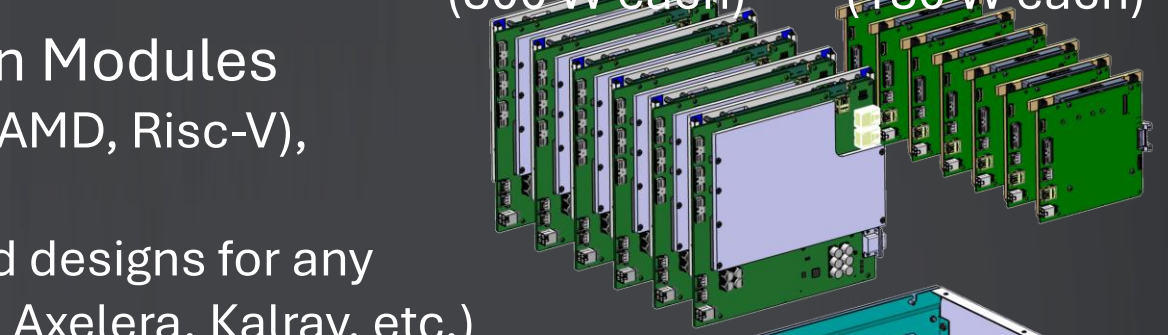


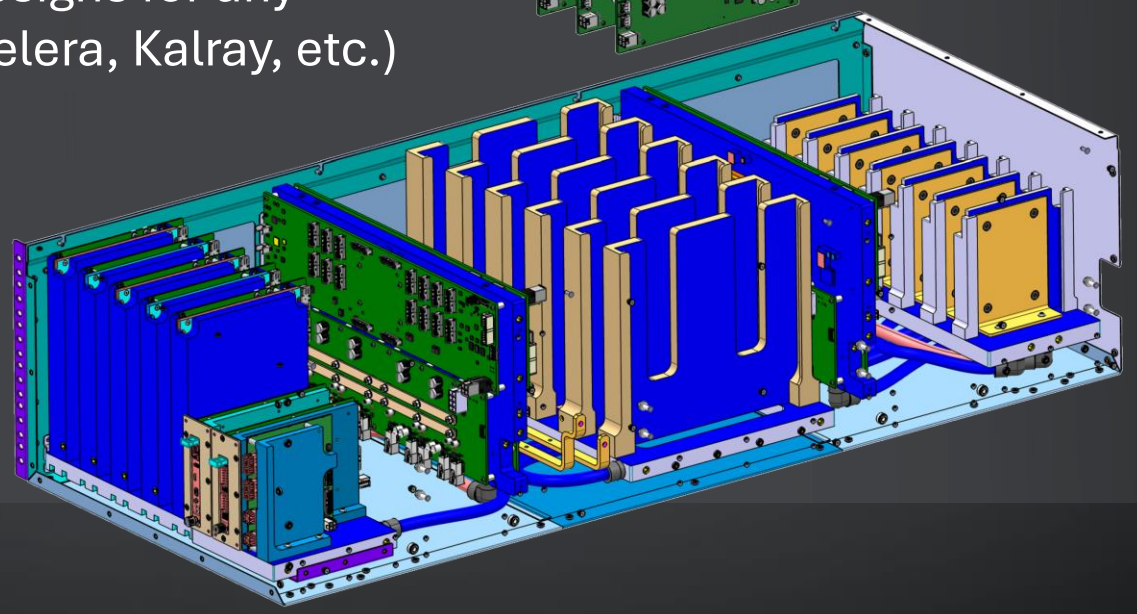
Custom CAPE Unit Configurations
Mix any Compute & Accelerator modules

Compute and Acceleration Modules

- Wide selection of CPU (Intel, AMD, Risc-V), FPGA (AMD), GPU (Nvidia)
- European OEMs custom board designs for any silicon incl. European SiPearl, Axelera, Kalray, etc.)

Industry Standard Modules
6x COM-HPC + 7x COM Express
(300 W each) (180 W each)

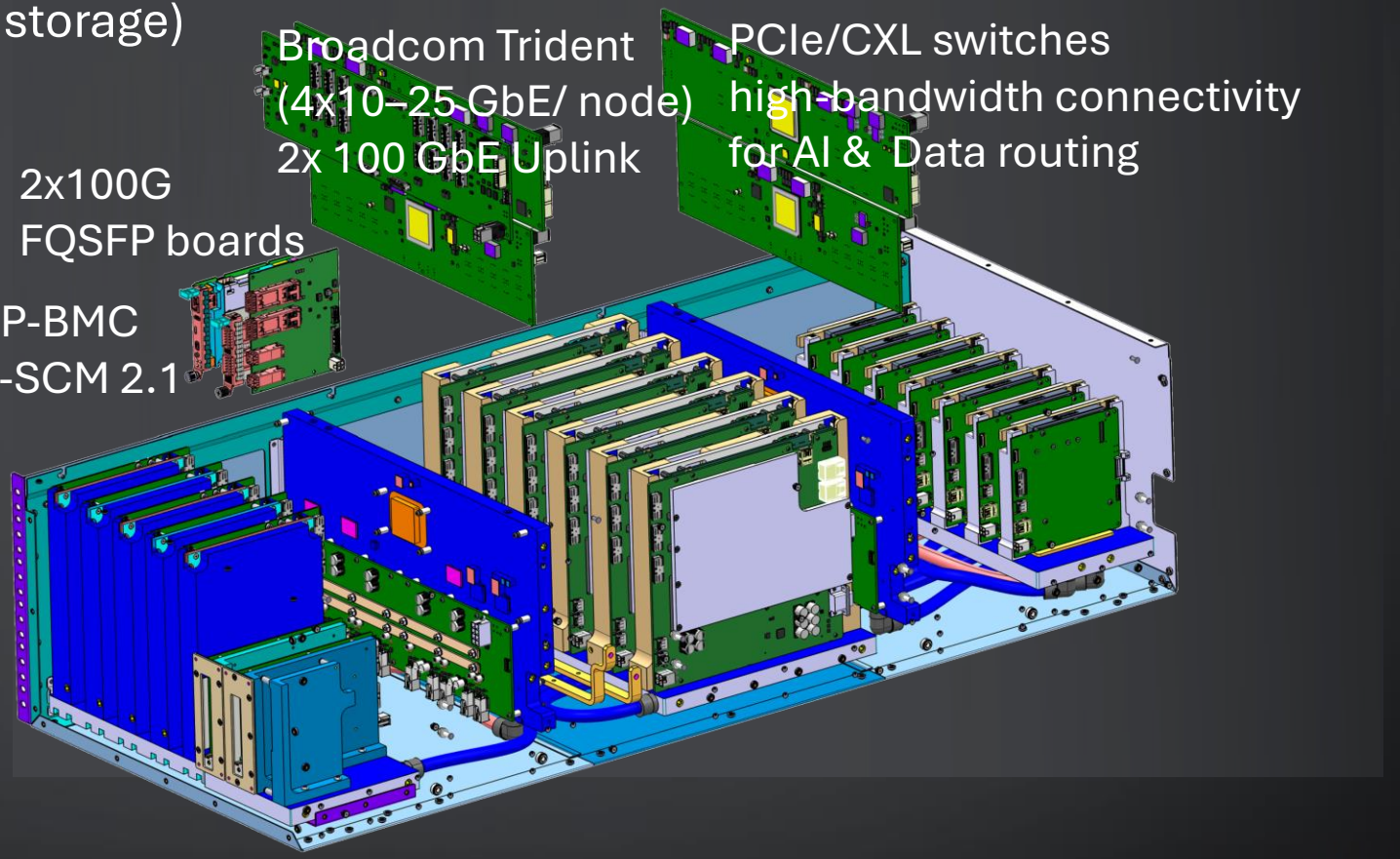




High Bandwidth dual fabric Ethernet, PCIe/CXL

CXL Server composability
from pools of Modules
(computation, accelerate, DRAM, storage)

High Bandwidth between
Local Data and AI

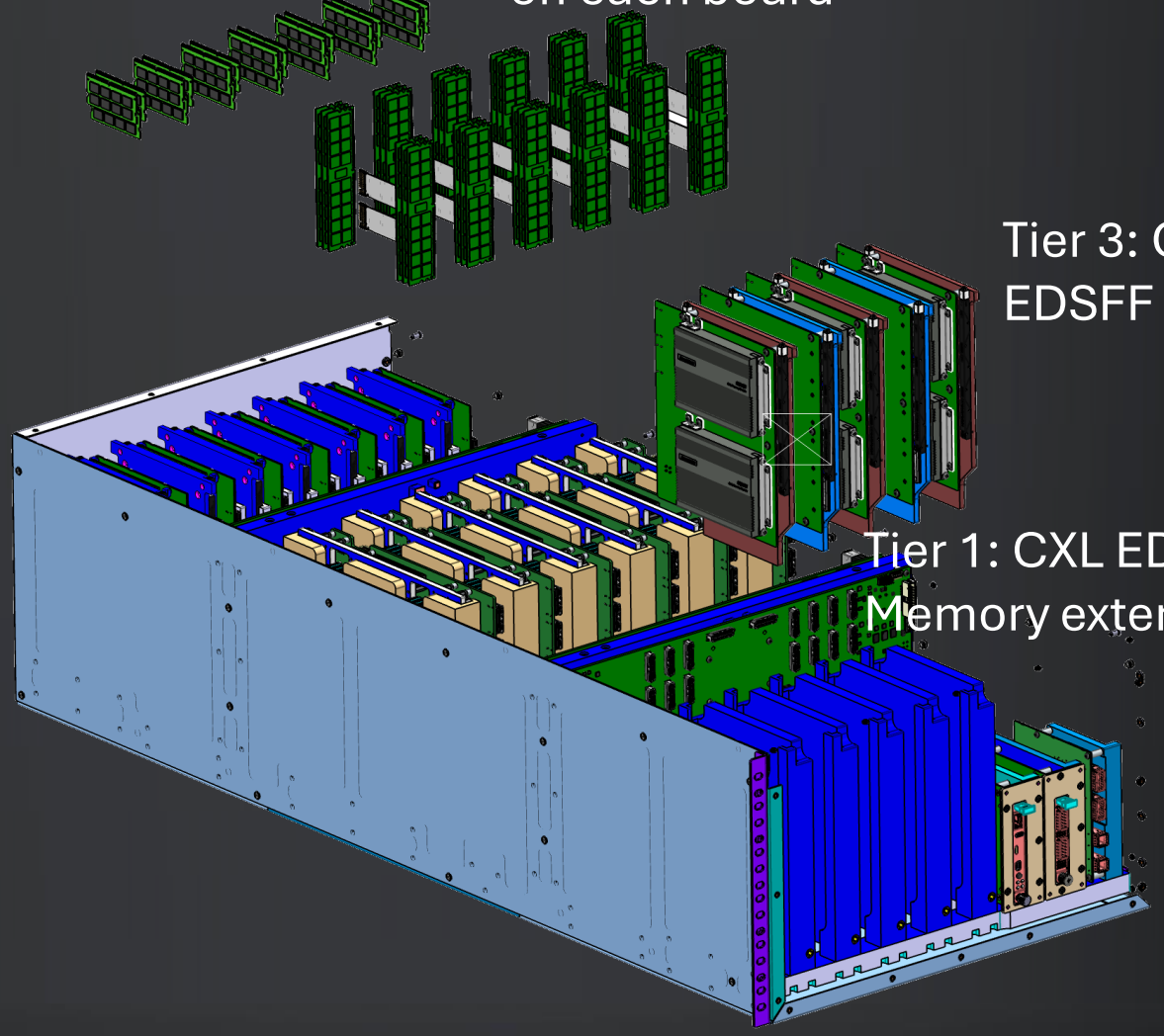


Storage Tiers

Tier 2: M2 NVMe 1-4TB
on each board

Tier 3: Over fabric
EDSFF E1.S SSDs

Tier 1: CXL EDSFF E3.S
Memory extension (Micron T.)

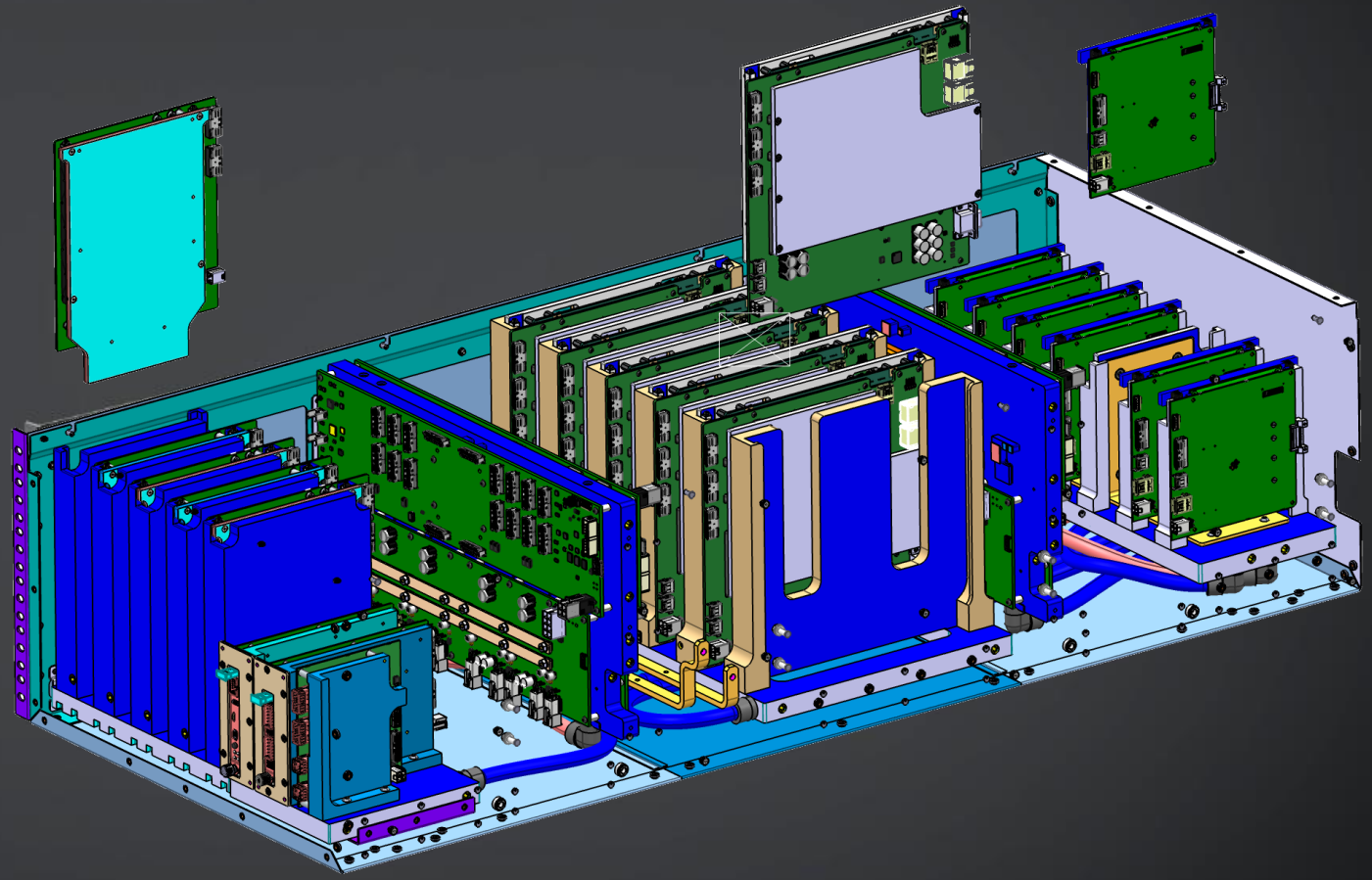


Composable hardware for AI and Big Data processing

Workload 2

Dual Fabric Ethernet & PCIe/CXL
Pools of CPU's, Accelerators, CXL Memory, Storage
At runtime composable server

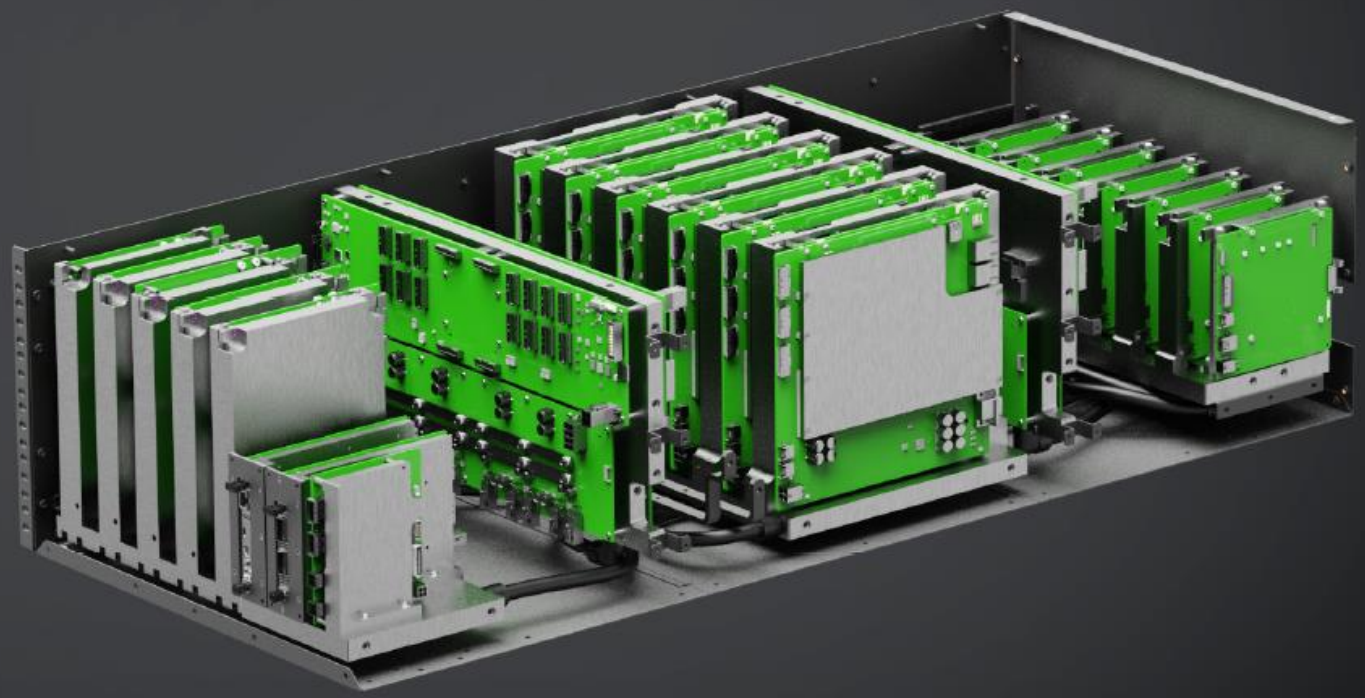
- # of CPU cores
- # of Accelerators
- # DRAM
- # Storage



Composable hardware for AI and Big Data processing

Dual Fabric Ethernet & PCIe/CXL
Pools of CPU’s, Accelerators, CXL Memory, Storage
At runtime composable server

- # of CPU cores
- # of Accelerators
- # DRAM
- # Storage

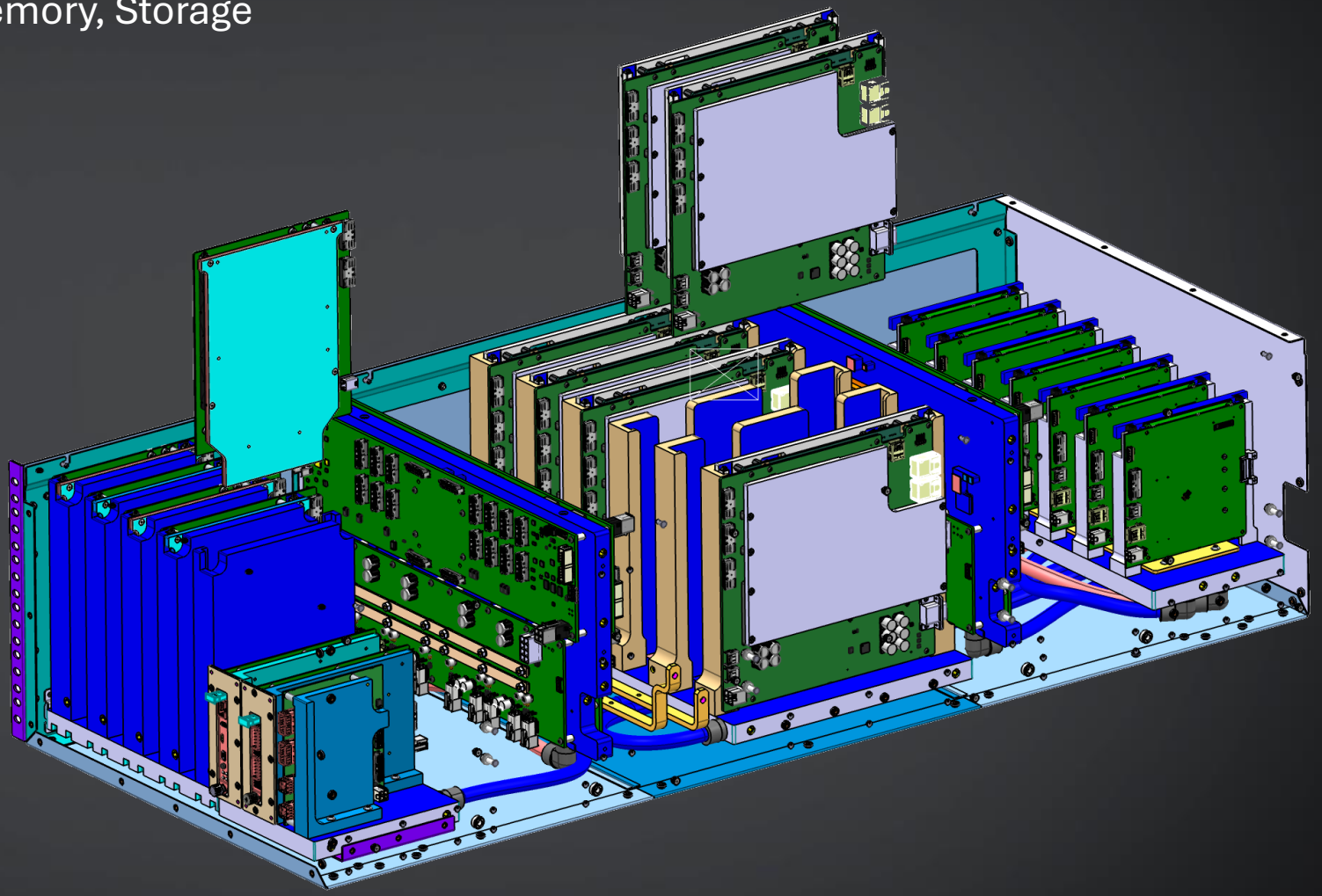


Composable hardware for AI and Big Data processing

Workload 3

Dual Fabric Ethernet & PCIe/CXL
Pools of CPU's, Accelerators, CXL Memory, Storage
At runtime composable server

- # of CPU cores
- # of Accelerators
- # DRAM
- # Storage



Composable hardware for AI and Big Data processing

Workload 1, 3

Dual Fabric Ethernet & PCIe/CXL
Pools of CPU's, Accelerators, CXL Memory, Storage
At runtime composable server

- # of CPU cores
- # of Accelerators
- # DRAM
- # Storage

