Campus to DataCenter Distributed AI Infrastructure

Maurizio Davini

Antonio Cisternino

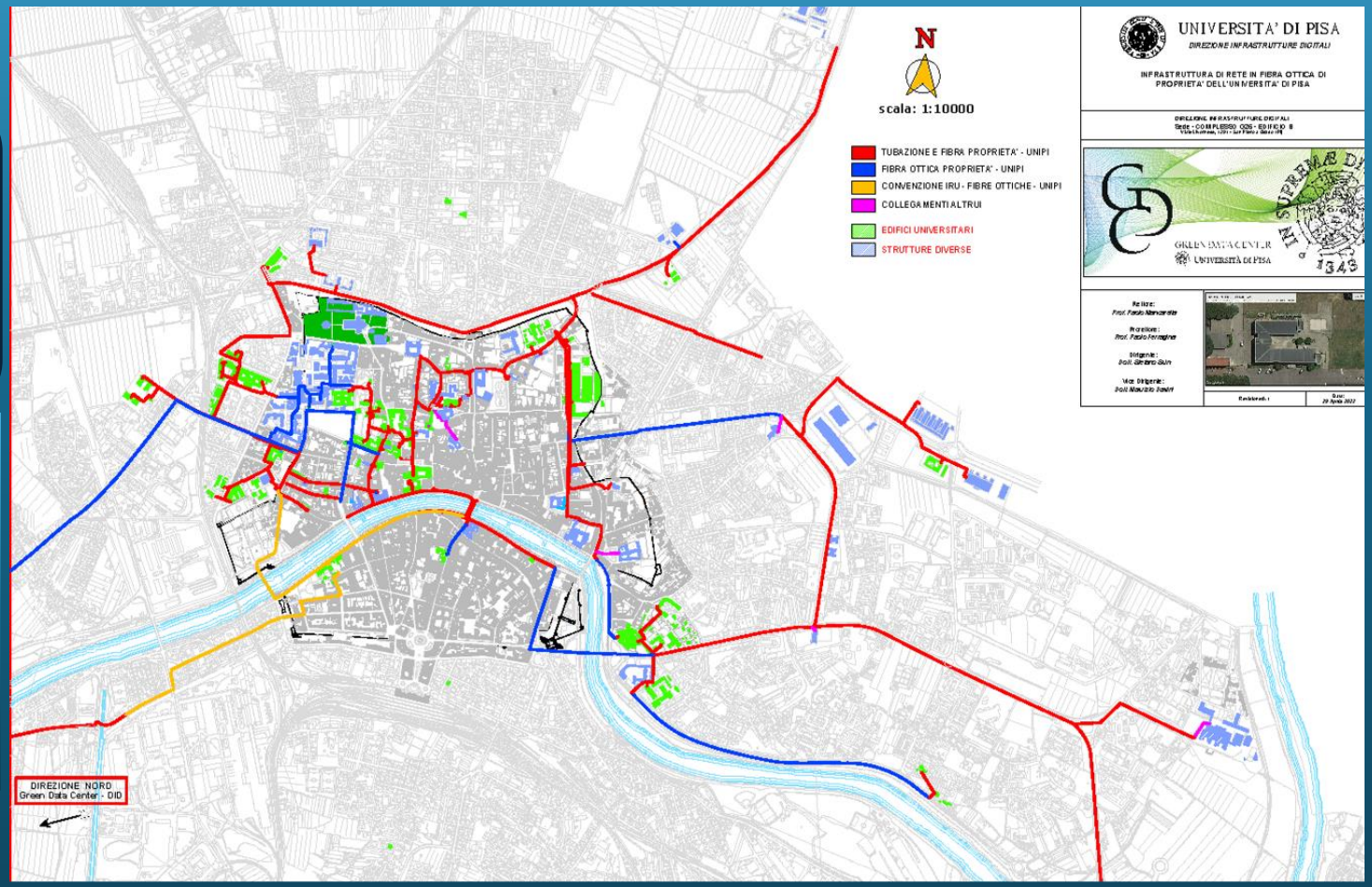
HPC-AI Swiss Conference 2026

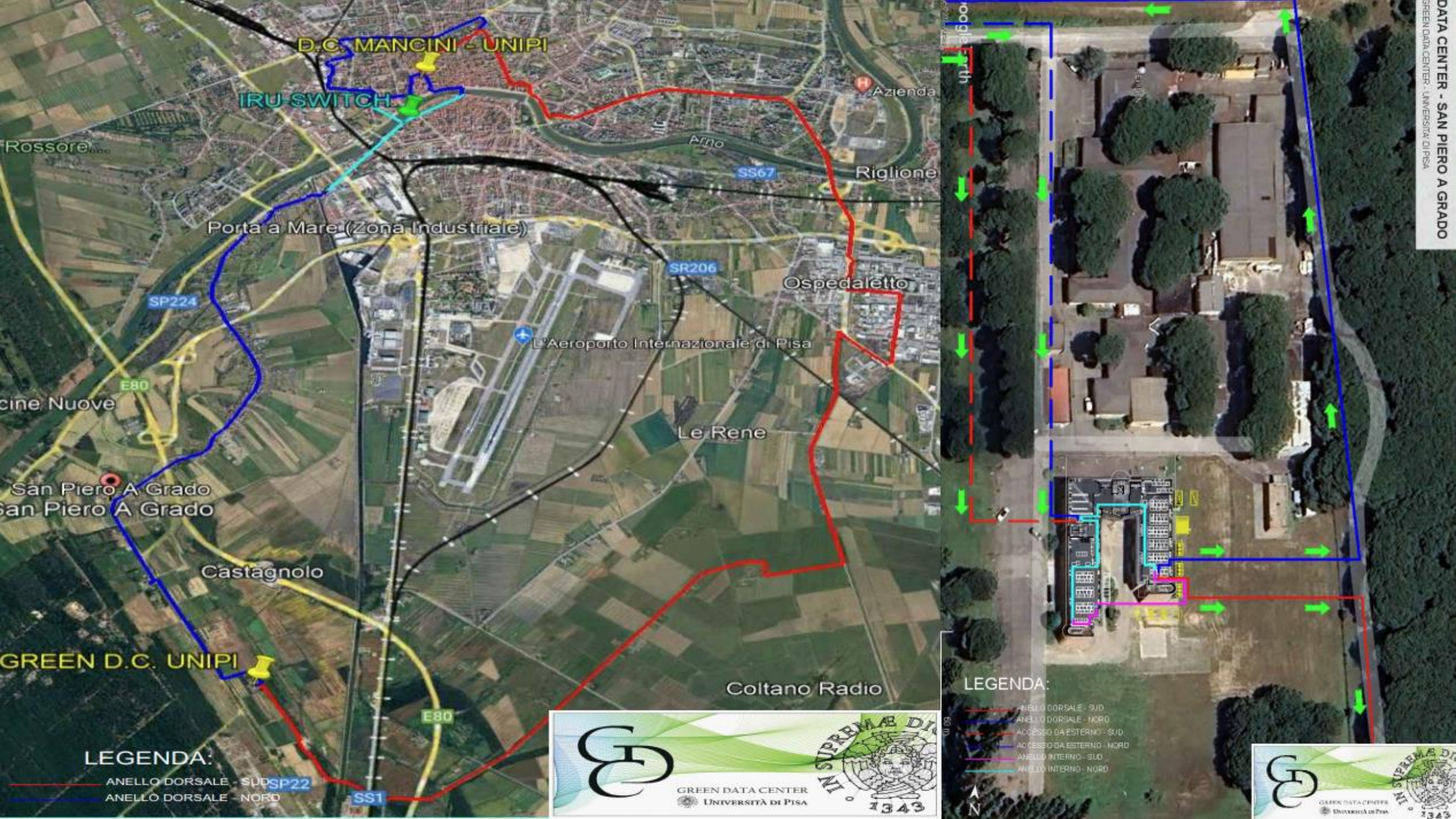
Campus Network

Some numbers

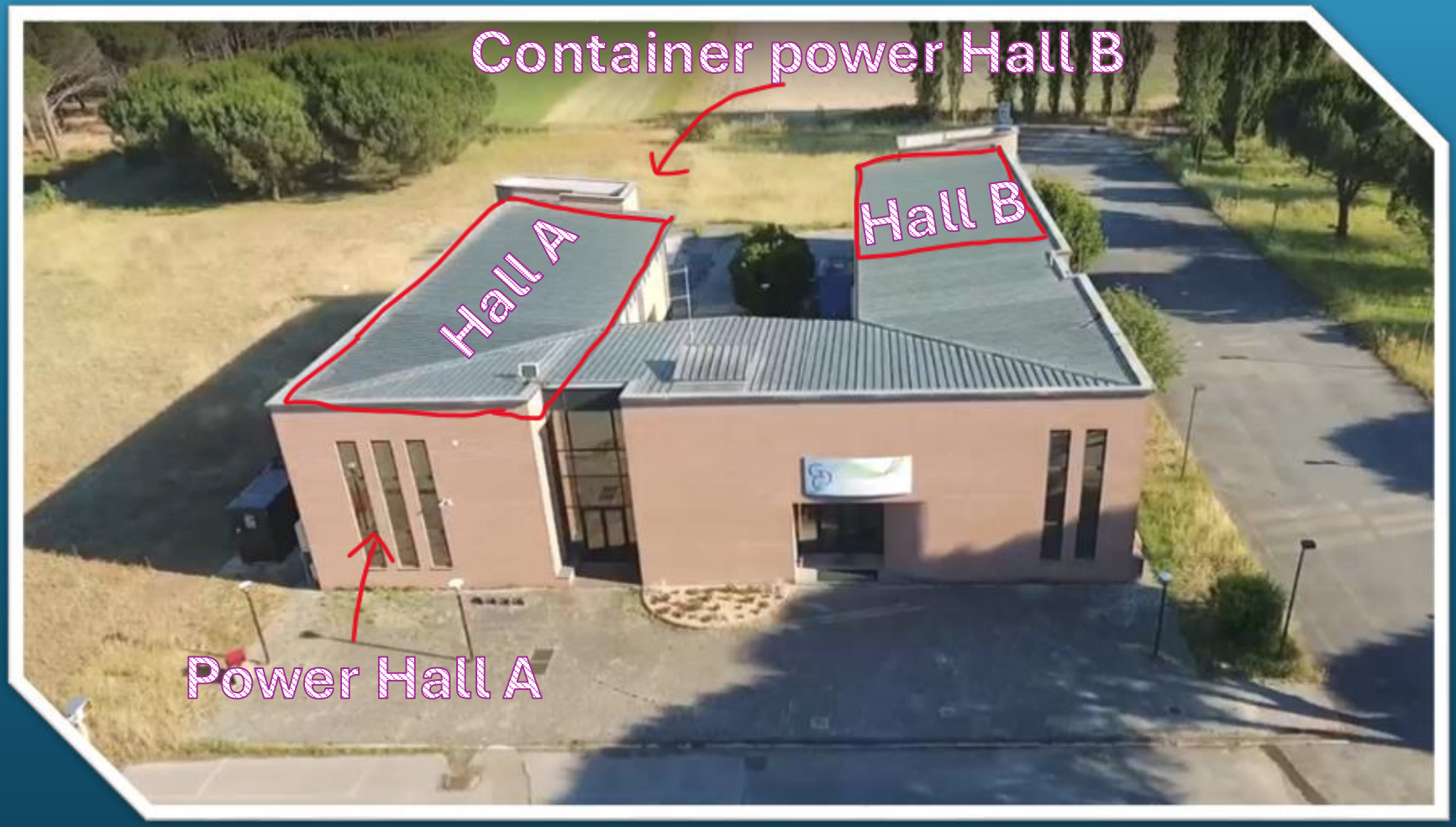
City wide
Campus

- 250 buildings
- 9.4k km SMF in 90km tubes

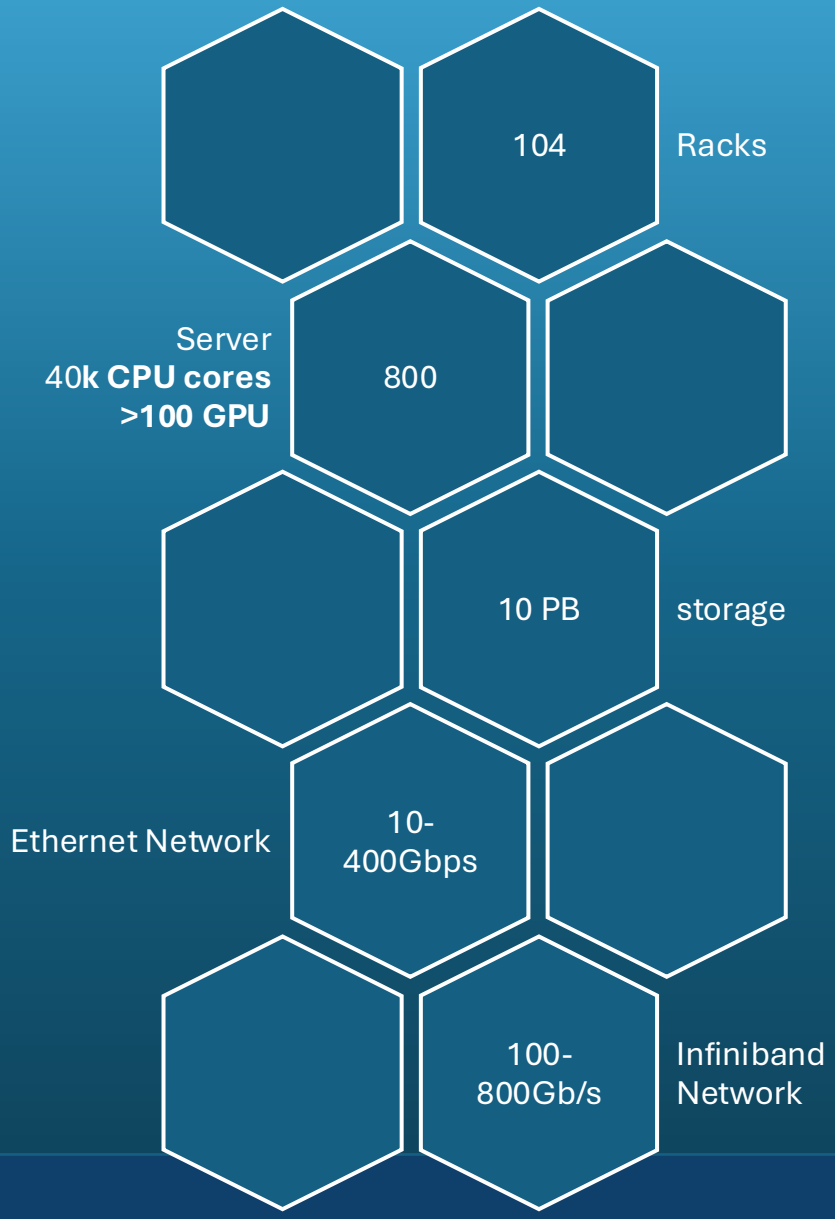



D.C. MANCINI - UNIPI
IRU-SWITCH
Rossore
Azienda
Arno
SS67
Riglione
Porta a Mare (Zona Industriale)
SR206
Ospedaletto
SP224
L'Aeroporto Internazionale di Pisa
E80
cine Nuove
Le Rene
San Piero A Grado
San Piero A Grado
Castagnolo
GREEN D.C. UNIPI
Coltano Radio
E80
SP22
SS1
LEGENDA:
ANELLO DORSALE - SUD
ANELLO DORSALE - NORD
GREEN DATA CENTER
UNIVERSITÀ DI PISA
DATA CENTER - SAN PIERO A GRADO
LEGENDA:
ANELLO DORSALE - SUD
ANELLO DORSALE - NORD
ACCESSO DA ESTERNO - SUD
ACCESSO DA ESTERNO - NORD
ANELLO INTERNO - SUD
ANELLO INTERNO - NORD
N

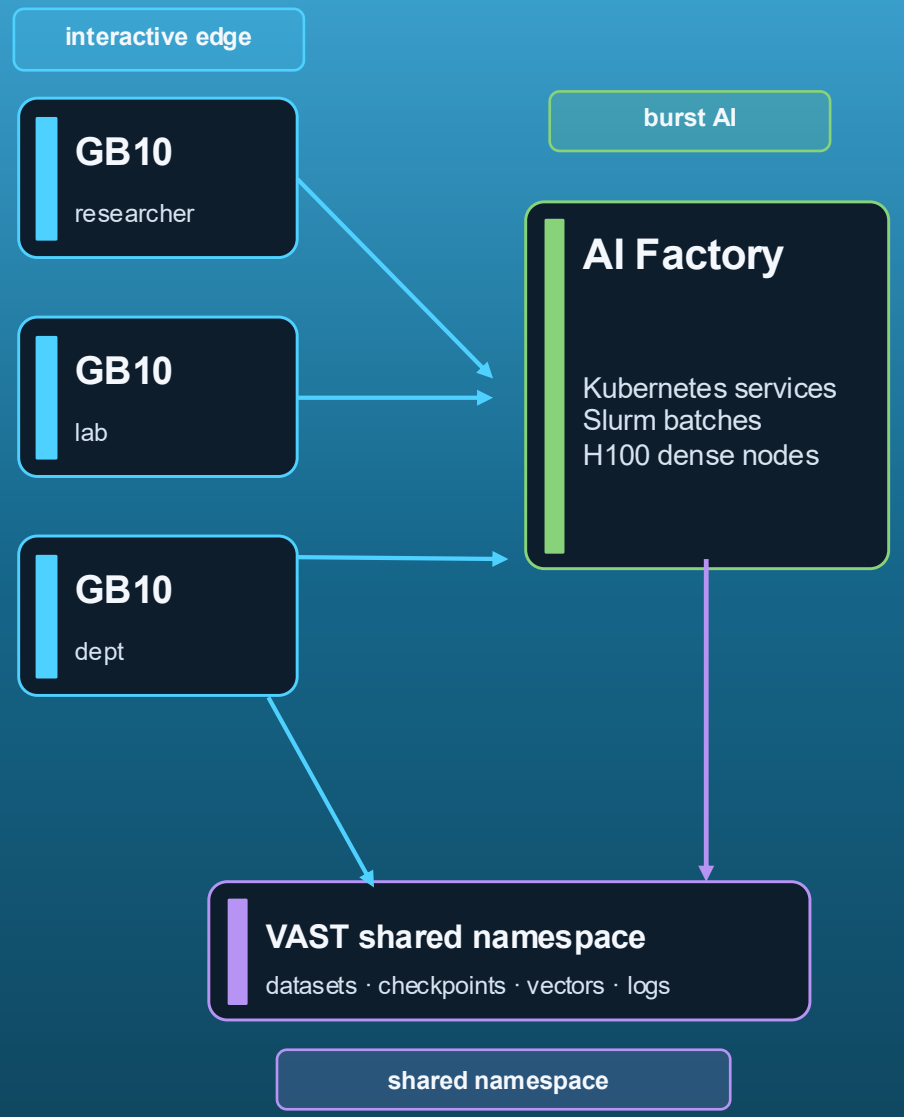
Unipi Green Datacenter



Some Numbers



The Model


interactive edge
GB10
researcher
GB10
lab
GB10
dept
burst AI
AI Factory
Kubernetes services
Slurm batches
H100 dense nodes
VAST shared namespace
datasets · checkpoints · vectors · logs
shared namespace

Campus-wide personal AI

Tier 0 · local

GB10 endpoints

- Personal AI with desk-side or room-side placement
- Departmental RAG, coding copilots, document assistants
- Agent runtime for OpenClaw / NemoClaw
- Low latency, local tools, local secrets, local memory

Tier 1 · burst

DataCenter AI factory

- Shared Kubernetes services and Slurm reservations
- Shared inference pools, embedding, evaluation, fine-tuning
- XE7745 for flexible PCIe scale-out; XE9640 and DGX H100 for dense scale-up
- Burst whenever memory, concurrency or topology become the real bottleneck

Tier 2 · shared

VAST unified namespace

- One namespace for raw data, curated corpora, vectors, checkpoints and logs
- Scale capacity independently from compute resources
- Avoid ad hoc copies between lab desks, HPC scratch and production services
- Enable data + models + metadata always-on for file/object consumers

KEEP LOCAL

human-loop / single-user

BURST/SERVE/TRAIN

batch + API multi-user

NFS / SMB / S3

data + models always-on

Design rule: keep the human loop local on GB10; use the DataCenter only once concurrency, memory, throughput or interconnect become dominant.

Hardware Roles and Operational Envelopes

Dell Pro Max with GB10

- up to 1 PFLOP FP4; 128 GB unified memory
- 20-core Arm CPU; DGX OS; local AI software stack
- 10GbE + ConnectX-7 for paired systems

PERSONAL AI / DESK-SIDE INFERENCE

Dell PowerEdge XE7745

- dual AMD 5th Gen EPYC; air-cooled 4U
- up to 8 double-wide 600W PCIe accelerators
- best for flexible shared inference and fine-tuning pools

FLEXIBLE POOL / SHARED ACCELERATION

Dell PowerEdge XE9640

- 4 × H100 SXM 700W with NVLink
- liquid-cooled 2U dense scale-up node
- best for tightly coupled training and HPC+AI

DENSE SCALE-UP / NVLINK

Nvidia DGX H100

- 8 × H100 Tensor Core GPUs; 640 GB total GPU memory
- 32 PFLOPS FP8; 4× NVSwitch; dual x86 host
- ConnectX-7 400 Gb/s InfiniBand or 200 Gb/s Ethernet fabric

REFERENCE / PREMIUM TRAINING PARTITION

Reference role	GB10	XE7745	XE9640	DGX H100
Workload class	human-loop	shared pool	dense scale-up	gold / reference
Scheduler	user-owned	K8s / Slurm	Slurm	Slurm / selected services
Best used for	local inferencing	embedding + APIs	all-reduce heavy	premium training

Interaction,Compute and Data Planes

Architectural goal: one simple user experience, shared data foundations, and clear service boundaries for operators.

Placement Policy

Match each workload to the right platform

decision flow

Placement decision

1. Dense GPU interconnect required?

- multi-GPU training
- all-reduce heavy jobs

→ **XE9640 / DGX**

2. Local, interactive, few users?

- restricted or departmental data
- model fits locally

→ **GB10**

3. Otherwise

- shared inference / embeddings
- higher concurrency or throughput

→ **XE7745**

workload map

Workload positioning

By latency sensitivity and compute/interconnect demand

Rule of thumb: keep small, interactive, data-local workloads on GB10; move shared or high-throughput services to XE7745; use XE9640 / DGX when GPU interconnect becomes critical.

Multi-arch by design

Standardize on OCI + shared APIs

	GB10 endpoint		DataCenter stack
UX / apps	OpenWebUI, notebooks, coding agents, departmental tools	same API	OpenWebUI, portals, APIs, notebooks
Model runtime	TensorRT-LLM, vLLM, Ollama, llama.cpp where suitable	same API	vLLM, Triton, NIM, PyTorch services
Data plane	NVMe cache, VAST mount, local vector DB	same API	VAST shared namespace, object + file data services
Containers	Docker / Podman, dev containers, OCI images	same API	Harbor / OCI registry, signed releases
Orchestration	single-node + light local schedulers	portable	Kubernetes + GPU Operator + Kueue/Volcano; Slurm + Pyxis/Enroot
APIs / drivers	CUDA where supported, model APIs, RAG services	portable	NCCL, PyTorch, Ray, RAPIDS/Spark, Dask
Base OS	DGX OS / Ubuntu Linux on Arm	portable	Ubuntu / RHEL on x86 + cluster toolchain

Critical discipline: package the same user-facing experience as OCI images for linux/arm64 and linux/amd64; never assume x86-only developer paths.

Campus fiber, RoCE in the pod, VAST on campus

1. Campus edge

Each department keeps a GB10 near the user for private, low-latency work.

Dept A → GB10 local AI

Dept B → GB10 local AI

Dept C → GB10 local AI

Fiber
private links
policy-aware QoS

Traffic QoS
interactive
batch
audit always on

2. AI pod

RoCE stays inside the pod. Put service, training, and dense GPU jobs where interconnect is strongest.

Control plane links
Kubernetes service traffic, Slurm paths, registry access, and observability stay inside the pod.

Data movement rule
Send references before bytes; stage large data only when topology or performance requires it.

3. Shared data on VAST

NFS / SMB / S3 for datasets, vectors, snapshots, checkpoints....

Architectural goal: keep AI close to the user, keep RoCE local to the pod, and keep the data shared on campus.

GB10 execution envelope

GB10 desk-side

1 PFLOP
FP4 COMPUTE

128 GB
UNIFIED MEMORY

10GbE
HOST NETWORK

20 Arm
CPU CORES

Practical read

Great for low-latency interaction, local context retention and secure tool access. Promote only the jobs that truly need multi-user scale or dense topology.

native local fit

private coding copilot · departmental RAG · VLM document triage · always-on agents on dedicated unit · interactive document assistants

works with quantization / expert mode

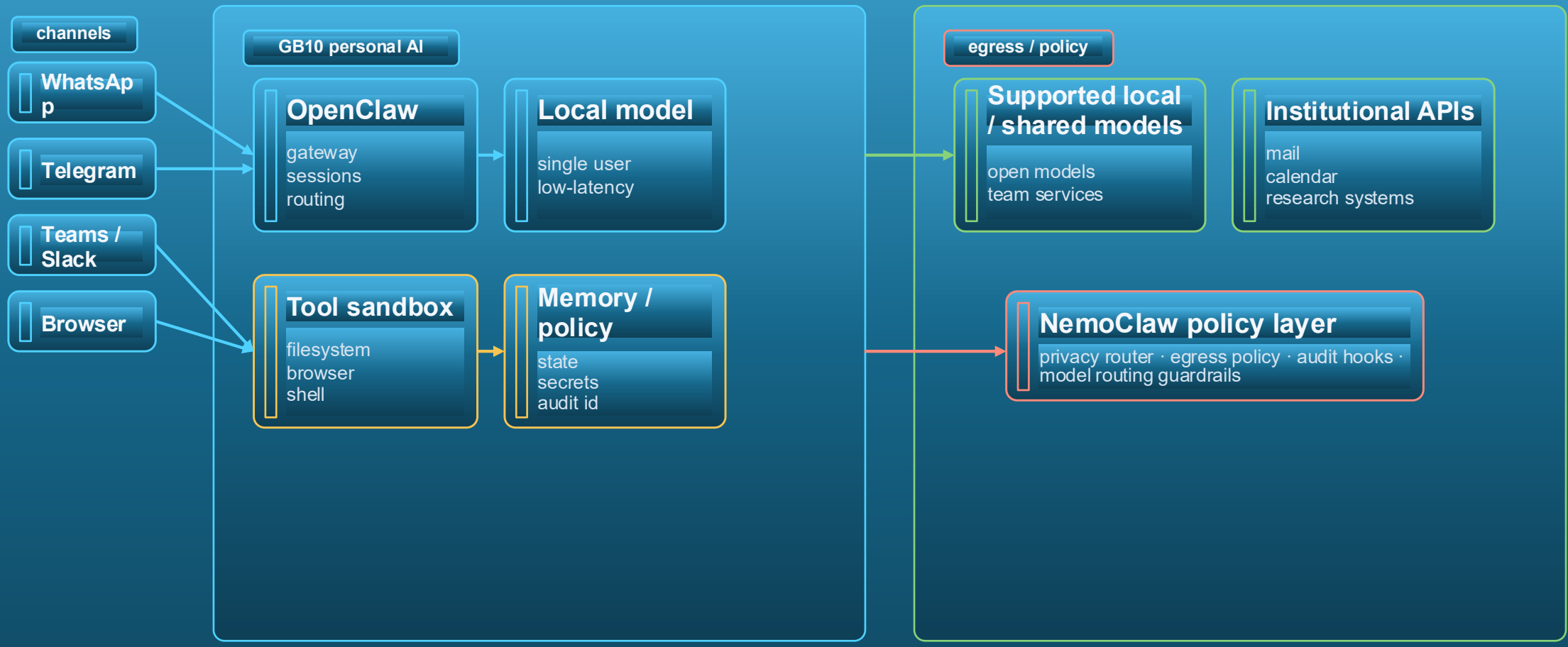
70B-class inferencing · tool-use agents · multimodal assistants · LoRA/QLoRA on small research sets · paired GB10 systems when local density matters

burst recommended

model serving for many users · dense fine-tuning · very large context / memory pressure · throughput-oriented embedding jobs · collective-heavy training

Move these workloads to XE7745 / XE9640 / DGX H100 when concurrency, memory or interconnect become the primary bottleneck.

Secure personal AI on GB10



Datacenter workload mapping

XE7745

flexible PCIe pool

- shared inference APIs and embedding jobs
- fine-tuning and evaluation sweeps
- best default burst tier for general services

SCHEDULER: K8S / SLURM

XE9640

dense 4-way NVLink node

- scale-up training and tightly coupled workloads
- better fit when local collectives dominate
- use for HPC+AI convergence and latency-sensitive scale-up

SCHEDULER: SLURM

DGX H100

gold / reference partition

- reference stack validation and premium training jobs
- topology-sensitive jobs and advanced debugging
- reserve for high-value work rather than generic shared inference

SCHEDULER: SLURM + SELECTED SERVICES

University workloads enabled by the fabric

Domain	Local on GB10	Central in DataCenter	Published outcome
CFD / FEM / HPC research	GB10 assistant for preprocessing, code generation, job steering, result interpretation	XE9640 / DGX H100 for surrogate training, coupled simulation+AI loops, large post-processing	team service with experiment memory, parameter search and reproducible reports
Biomedical / clinical research	private document triage, coding copilot, controlled local retrieval on sensitive material	XE7745 shared inference for de-identified analysis, embeddings and evaluation	auditable service with approval path and restricted namespace
Digital humanities	local OCR cleanup, transcription aid, multilingual RAG, curation tools	XE7745 burst for collection-scale indexing, multimodal inference and batch annotation	institutional research portal with persistent search and provenance
Teaching and lab copilots	course-local assistants on GB10 for exercises and lab support	central shared models for peak semester load and analytics	departmental assistant with controlled rollout and availability targets

Same physical fabric, different semantic contracts: personal assistance, team infrastructure and institutional services coexist without flattening requirements.

A campus-owned fiber fabric makes it possible to place AI where it belongs: near people when latency and privacy matter, in the DataCenter when scale and shared utilization dominate.

1

P1-C campusness becomes useful

Researchers can build, test and refine models locally without turning the workflow into remote-only computing.

2

Central GPUs are used when justified

GB10 absorbs personal interaction loops; shared pools handle throughput, burst, service APIs and dense training.

3

Reproducibility and governance are built in

One namespace, signed artifacts and auditable placement decisions enable institutional operation.

Next technical work: topology-aware scheduler, memory-aware placement, burst policies, and federation with national or EuroHPC resources.


Questions