pacewisdom

WHITE PAPER

Sovereign Clinical AI at the Edge

A reference stack on Wind River Cloud Platform

In collaboration with

WNDRVR

Sovereign Clinical AI at the Edge

A reference stack on Wind River Cloud Platform

Inference on identifiable patient and trial data belongs on private infrastructure at the edge. The regulatory, contractual and cost arguments all point there, and the organisations we work with (pharmaceutical companies, CROs and national medical agencies) have already concluded it. Their contracts with upstream sponsors say the data does not leave the premises; their regulators say the same; and owned inference is a fixed cost that amortises across sites.

The open question is operational. An IT team has to run GPU inference, agent orchestration, databases and regulated legacy applications across many sites, usually without a platform team to do it.

The reference stack in this paper has eight workload classes, each with a stated resource profile, state model and container-or-VM placement, mapped onto Wind River Cloud Platform (WRCP). WRCP is the southbound layer: a StarlingX-based private cloud that runs containers and virtual machines side by side on commodity x86, with Ceph storage, GPU scheduling, and a central-cloud-plus-sub-clouds model that lets one team operate dozens of sites from one console. Pace Wisdom is the northbound layer: the clinical AI workloads, the integration and the deployment practice.

WRCP's distributed topology matches how clinical AI gets deployed: a central private cloud where models are trained and validated on de-identified data, and site sub-clouds where inference runs on live patient data and nothing leaves.

Section 8 has a parametric five-year cost-of-ownership model for the stack, by site tier. The inputs are exposed, so your own volumes and rates decide the answer.

HIGHLIGHTS

Inference on identifiable patient and trial data belongs on **private infrastructure at the edge**. Sponsor contracts, regulators and cost all point there.

Eight workload classes, each with a stated **container-or-VM placement**, mapped onto a private cloud that runs both on one node.

One **shared operator** makes the sub-cloud the cheapest site to own: about a third of a stand-alone cluster over five years.

Disclaimer. This white paper is a proposed reference architecture rather than a validated WRCP deployment, particularly for the AI inference and GPU configurations. It is a work in progress, not a validated solution implementation.

1. Why private cloud at the edge

1.1 The three pillars

Sovereignty is three separate guarantees, and only owned infrastructure inside the perimeter gives all three.

- **Data sovereignty**: clinical records, DICOM, safety reports, prompts, embeddings and model weights never cross the perimeter.
- **Operational sovereignty**: runtime, keys, patches and hypervisor are administered by the enterprise's own staff. No foreign subpoena reaches a server the provider cannot see.
- **Technological sovereignty**: open-weight models on hardware the enterprise owns cannot be deprecated, repriced or geofenced by a vendor.

Hyperscaler "sovereign" offerings deliver residency, which is a location claim, not control. A private-link path to a managed inference endpoint does not move the provider's fleet into your data centre; data that is in-region on a foreign-owned cloud remains reachable under that provider's home jurisdiction.

1.2 The data-class rule

Data class	Examples	Where inference runs
Identifiable PHI / trial subject data	EHR text, DICOM, ICSRs, case narratives, consent forms	**Edge sub-cloud, inside the site perimeter**
Pseudonymised, re-identifiable in-house	Coded trial datasets, linked registries	**Edge sub-cloud or central private cloud**
De-identified, non-reversible	Anonymised corpora for fine-tuning, aggregate analytics	Central private cloud; public cloud permissible for training only
Non-personal	SOPs, protocols, literature	Anywhere; edge is usually cheaper at volume

The rule in one line: **identifiable data in production is inferred at the edge; de-identified data may be trained on centrally; nothing else leaves.**

1.3 What the regulators demand

Regime	Demand	Edge answer
HIPAA / GDPR	Transfer and sub-processor obligations for any third-party processing	No transfer, no sub-processor, no residency attestation
India DPDP Act and Rules (notified 13 Nov 2025)	Consent-driven; localisation with conditional cross-border transfer; Significant Data Fiduciaries carry enhanced duties; government may restrict specified categories	Large hospital groups, CROs and pharma should assume SDF designation and that health data will be a restricted category
EU AI Act, high-risk tier	Stand-alone high-risk systems from 2 Dec 2027; AI in regulated products incl. medical devices from 2 Aug 2028; lifecycle risk management, human oversight, append-only logs. Dates per the 2026 Digital Omnibus, subject to its final text and harmonised-standard timelines	Continuous logging and oversight are evidenced on infrastructure you control

> Naming the workload classes is what makes the platform conversation tractable: each class has a resource profile, a state model, and a natural placement as a container or a VM.

2. Workload classes in a clinical AI stack

Every clinical AI deployment we have built decomposes into the same eight workload classes. Naming them is what makes the platform conversation tractable: each class has a resource profile, a state model, and a natural placement as a container or a VM.

Legacy applications are why the platform has to run VMs at all. Every pharma site has a validated system that will not be containerised in this decade.

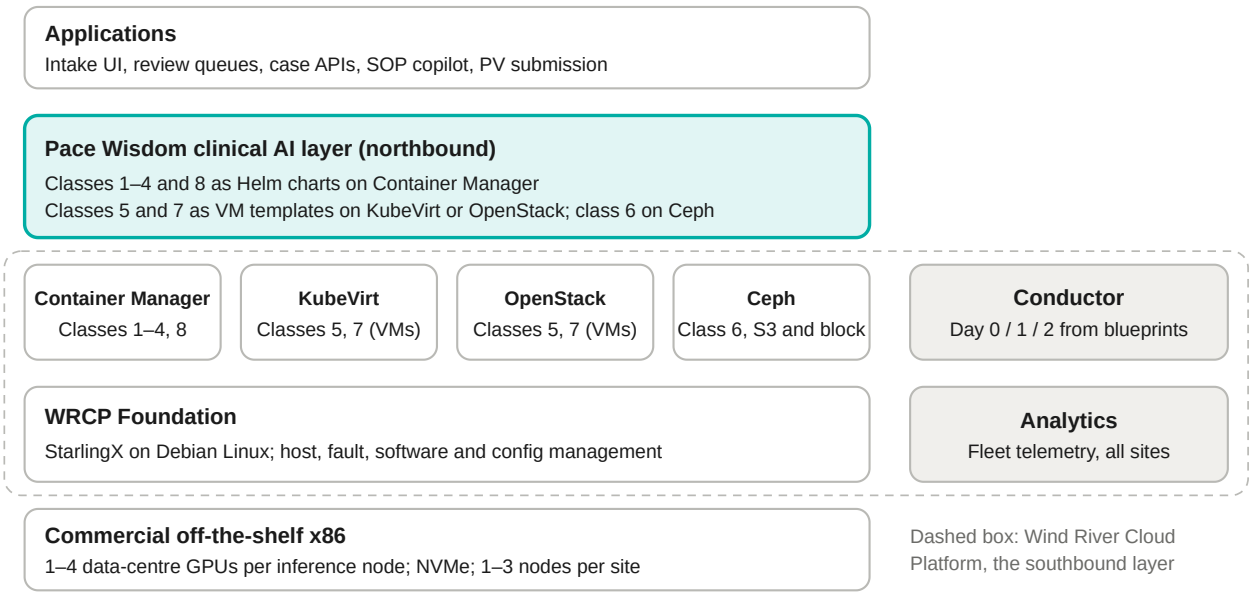
#	Class	What it is	Runs as	Resource profile	State
1	**Business logic & applications**	Intake UI, review queues, case APIs, reporting	Containers	CPU, 2–8 vCPU, stateless	None (state in DB)
2	**Workflow & message orchestration**	n8n-class workflow engine, message queue, scheduler	Containers	CPU, 4–8 vCPU	Durable queue on persistent volume
3	**AI inference**	vLLM-class LLM serving, vision-language models, medical ASR, embedding models	Containers with GPU via device plugin; VM with GPU passthrough where driver isolation or a vendor-certified image is required	GPU-bound; 1 GPU per model instance; 24–80 GB VRAM by model	Model cache on shared storage
4	**Agent orchestration & gateway**	LangGraph-class agent runtime; LiteLLM-class gateway (OpenAI-compatible API, routing, fallback, rate limits)	Containers	CPU, 4–8 vCPU; memory for context	Prompt and completion log, append-only

#	Class	What it is	Runs as	Resource profile	State
5	**Database**	Postgres with pgvector; audit store	VM, or stateful container with a Ceph-backed PVC	Memory and IOPS; 16–64 GB RAM	Persistent, backed up
6	**Object & model storage**	Documents, DICOM, dictation audio, model weights	Ceph RGW (S3-compatible) and RBD, native to WRCP	Capacity; NVMe tier for model cache	Persistent, replicated
7	**Legacy & regulated applications**	EDC, pharmacovigilance systems, Windows-bound tools, validated vendor appliances	VMs via KubeVirt or Wind River OpenStack	Varies; typically 4–16 vCPU	Persistent; often change-controlled
8	**Observability & governance**	Logging, metrics, RBAC/OIDC, attestation, audit export	Containers, plus WRCP Analytics for infrastructure telemetry	CPU, small	Append-only

Three placement notes that come up in every design review:

- **Inference as container by default.** The GPU is scheduled through the Kubernetes device plugin, which is what StarlingX supports for NVIDIA via the GPU Operator. A VM with PCI passthrough is the fallback when a model vendor ships a certified VM image or when a workload needs a different driver stack from its neighbours.
- **Database as VM unless the team already runs stateful Kubernetes in production.** Both work on WRCP; the VM is the lower-risk default for a team whose DBA practice is VM-shaped.
- **Legacy apps are why the platform has to run VMs at all.** Every pharma site has a validated system that will not be containerised in this decade. A platform that only runs containers forces a second platform; WRCP runs both on one node.

3. The reference stack on Wind River Cloud Platform



Reference stack on WRCP

Everything inside the dashed box is one WRCP install on one ISO; the accented layer is what Pace Wisdom deploys onto it. Conductor and Analytics operate the whole fleet from the central cloud.

Hardware. Commercial off-the-shelf x86 servers. For inference nodes, one to four data-centre GPUs per server (L40S-class for most inference; H100-class where model size demands it) or a single workstation-class 48 GB card for the lightest sites. Ceph replicates across three nodes at a site. A single all-in-one node runs Ceph at replication factor 1, which keeps the S3 and block APIs but gives no host or drive fault tolerance; a two-node (AIO-DX) sub-cloud is the smallest configuration with replicated Ceph.

WRCP Foundation. StarlingX-derived, Debian-based low-latency Linux, hardened and commercially supported by Wind River, delivered as one ISO. Kubernetes for containers; KubeVirt and containerised OpenStack for VMs, both on the same control plane; Ceph software-defined storage; host, fault, software and configuration management built in. Node-based licensing, not per-core or per-VM.

Clinical AI layer. The eight workload classes from Section 2, deployed as Helm charts (containers) and VM templates (KubeVirt/OpenStack). The gateway exposes one OpenAI-compatible endpoint per site, so application code is identical whether it runs against a site sub-cloud, the central cloud, or, for de-identified development, a public cloud.

Conductor. Day 0 provisioning of a new site from blueprint; day 1 deployment of the clinical AI layer; day 2 upgrades, scaling and configuration audit across every site from the central cloud. Cloudify-based, blueprint-driven, with plugins for public clouds as well: the same tool that pushes a model to twenty sub-clouds can stage a training job on a hyperscaler.

Analytics. Infrastructure telemetry across central and sub-clouds. The clinical AI layer's own observability (class 8) feeds application metrics alongside it.

> Signed model artefacts and configuration move outward; telemetry and audit summaries move inward. Patient data never moves.

4. Topologies

WRCP's distributed cloud model is a central cloud (the System Controller) managing a pool of autonomous sub-clouds, each of which can be a single all-in-one node, a two-node HA pair, or a standard multi-node cluster. Three topologies cover the clinical cases we see.

4.1 Central private cloud + N site sub-clouds

The hospital group, the multi-site CRO, the pharma company with several plants and a PV function.

- **Central cloud:** model training and fine-tuning on de-identified data; evaluation; the model registry; Conductor and Analytics; the audit aggregation point. Multi-node with larger GPUs where training is in scope; a single control node where it is not.
- **Site sub-clouds:** inference on live data; the local gateway, agent runtime, workflow engine, database and object store; local legacy VMs. One to three nodes each.
- **What moves between them:** signed model artefacts and configuration outward; telemetry and audit summaries inward. Patient data never moves.

This is the "train centrally, infer at the site" pattern made operational. A sub-cloud keeps running if the link to the centre drops; Conductor reconciles when it returns.

4.2 Single-site air-gapped

The medical agency, the regulator, the manufacturing plant with a hard no-network policy.

One standard cluster, no System Controller link, no external route at all. Models, platform updates and blueprints arrive on signed media and are applied through Conductor running locally. Ceph provides the object store; everything in Section 2 runs on the one cluster. This is the strictest configuration and the one with the fewest moving parts.

A sub-cloud keeps running if the link to the centre drops. Conductor reconciles when it returns.

4.3 Single workstation-tier node

The trial site, the small clinic, the departmental pilot.

One all-in-one WRCP node with a single workstation GPU, local storage, and the lightweight classes only: gateway, a small model, the SOP corpus, a local database. Managed as a sub-cloud of the central cloud where one exists; stand-alone where it does not.

Identifiable data in production is inferred at the edge. De-identified data may be trained on centrally. Nothing else leaves.

5. Three deployments on this stack

Composite descriptions drawn from live Pace Wisdom engagements. Outcomes are indicative; client names are withheld.

5.1 Vision-language intake for pharmacovigilance

A PV function receiving adverse-event reports in dozens of layouts, languages and scan qualities, with a 15-day regulatory clock on serious cases. A small vision-language model reads each document image directly and emits a structured case record; a reviewer confirms before submission.

On this stack: topology 4.1. The model is fine-tuned centrally on de-identified historical cases, promoted by Conductor to PV-site sub-clouds. Classes 1, 2, 3, 4, 5, 6, 8 at the site; class 7 for the validated PV system the records are submitted into. Two L40S-class GPUs per site.

Indicative outcome: structured-field accuracy in the high 90s versus low 80s for the OCR-plus-rules baseline; manual-correction rate more than halved; per-document cost on owned hardware in the cents at the function's volume.

5.2 Air-gapped narrative generation and anonymisation

A medical agency drafting case narratives and producing anonymised document packages for publication, from source documents that are the most sensitive artefacts the organisation holds.

On this stack: topology 4.2. A mid-size open-weight model (70B class, quantised to INT4 or FP8 so it fits two 48 GB cards) on a multi-GPU node drafts narratives from structured case data; a separate pipeline does entity detection and redaction with tiered review. The air gap means the anonymisation process itself cannot leak. Classes 1–6 and 8 on one cluster; updates by signed media.

Indicative outcome: narrative drafting from hours to minutes per case with a medical writer in the loop; anonymisation throughput up several-fold; zero external data movement by construction.

5.3 SOP copilot at the trial site

Coordinators and assistants at research sites following dozens of protocol- and site-specific SOPs, with deviations carrying regulatory consequence, on unreliable connectivity.

On this stack: topology 4.3. SOPs and protocols (non-personal data) embedded and served through retrieval-augmented generation on a

small general-purpose model with structured output. The copilot cites the governing document and never sees patient data. Classes 1, 3, 4, 5 on a single node.

Indicative outcome: protocol questions answered in seconds with citation; fewer deviations attributable to SOP misreading; near-zero marginal cost per query; works offline.

> The second site is where edge programmes succeed or stall. The blueprint is the deliverable that matters.

6. Operations

Day 0. A new site is a Conductor blueprint: node count, GPU inventory, storage layout, network segments. Zero-touch provisioning from the central cloud; the site team racks and cables.

Day 1. The clinical AI layer deploys as a second blueprint: Helm charts for classes 1–4 and 8, VM templates for 5 and 7, Ceph pools for 6. One blueprint, N sites.

Day 2. Model promotion is a signed artefact pushed from the central registry to selected sub-clouds, with rollback. Platform upgrades run through WRCP's software management on a maintenance window per sub-cloud. Configuration audit runs centrally and reports drift per site, the single largest operational risk in any edge programme. Model artefacts run 15–40 GB, so each site mirrors the registry and pre-fetches the next release in the background before cutover; the WAN never gates a release.

Air-gap path. For topology 4.2, the same blueprints and artefacts are exported to signed media and imported by a local Conductor instance. Nothing in the design assumes a network route exists.

GPU scheduling. Inference containers request `nvidia.com/gpu` through the device plugin; the scheduler places them on GPU nodes. Where a VM needs the GPU, KubeVirt passes the PCI device through. A site with one GPU serves one model at a time well; a site that needs two models concurrently needs two GPUs or a MIG-capable card.

Failure modes we design for. WAN loss (sub-cloud keeps serving); node loss (two-node HA for sites that cannot tolerate an outage); GPU loss (gateway falls back to a smaller CPU-capable model with a degraded-mode flag in the audit log); model regression (rollback to the previous signed artefact).

Configuration drift between sites is the single largest operational risk in any edge programme. Audit runs centrally and reports it per site.

7. Governance and human oversight

The platform makes safety evidenceable; the workflow has to enforce it.

Tier	Example workloads	Autonomy	Oversight
Low	SOP search, policy Q&A, administrative drafting on non-PHI	Autonomous; logged	Periodic spot-check
Medium	De-identification, case-intake extraction, narrative drafts, coding support	Semi-autonomous; outputs staged for review	Mandatory qualified review before record commit
High	Diagnostic interpretation, clinical decision support, treatment planning	Advisory only	Mandatory physician sign-off with immutable audit

Open medical models are developer research tools, not medical devices; using them for direct diagnosis or treatment requires device approval in the relevant jurisdiction. Local clinical validation on de-identified historical data precedes any medium- or high-tier deployment. Every prompt, completion, model version and operator action lands in the append-only log (class 8), exported to the central cloud where one exists. For GxP-adjacent use the same log is the basis for 21 CFR Part 11 and EU Annex 11 compliance on electronic records: attributable, time-stamped, tamper-evident, with electronic signatures at the review checkpoints.

8. Total cost of ownership

What it costs to own and run the stack, per site, over 5 years. The inputs are figures an IT team already has; change them and the outputs follow. The WRCP subscription and integration services are commercial quotes and are carried as placeholders.

8.1 Inputs

Input	Value	Basis
Site volume (PV intake scenario)	2,000 pages and 12 M tokens per site per working day; 250 days	§5.1
Workstation node	1 node, 1× 48 GB workstation GPU, local storage	$15,000 installed; 0.6 kW
Site cluster	3 nodes: 2 inference nodes with 2× L40S-class GPUs each, 1 control and storage node	$95,000 installed; 3.0 kW
Sub-cloud node	1 node, 2× L40S-class GPUs; single-replica Ceph, no host fault tolerance (a 2-node HA sub-cloud is the step up)	$40,000 installed; 1.2 kW

Input	Value	Basis
Central control and staging node (shared)	1 node, 2× L40S-class GPUs; System Controller, model registry, Conductor, Analytics; LoRA-class fine-tuning only, heavy training out of scope	$40,000 installed; 1.2 kW; shared by N sites
Electricity	$0.10 per kWh effective facility rate including cooling; 8,760 h per year	Assumption
Platform operations	$80,000 per FTE-year	Assumption; §8.4 gives an India rate
Ops effort	Workstation node 0.1 FTE; site cluster 0.5 FTE; central plus sub-clouds 0.5 FTE shared across N sites	Assumption
WRCP subscription	per node	Placeholder for Wind River pricing
Integration and support	per engagement	Placeholder

Hardware basis: L40S-class cards price well below H100 ($22–30k PCIe); $40,000 per 2-GPU node is an assumption, not a quote.

8.2 Five-year cost per site, by tier

N = 4 sites for the sub-cloud tier.

Cost line	Workstation node (§4.3)	Single-site cluster (§4.2)	Sub-cloud of a central cloud (§4.1)
Hardware	$15,000	$95,000	$50,000 (node plus a quarter of the central node)
Power and cooling	$2,600	$13,000	$6,600
Platform operations	$40,000 (0.1 FTE)	$200,000 (0.5 FTE)	$50,000 (an eighth of an FTE)
WRCP subscription	placeholder	placeholder	placeholder
Integration and support	placeholder	placeholder	placeholder
Five-year total before placeholders	**$57,600**	**$308,000**	**$106,600**
Annual run rate after year 0	$8,500	$42,600	$11,300



Cumulative cost per site over 5 years, before WRCP subscription and integration services · computed from the §8.1 inputs

Operations, not hardware, is the largest line in every tier except the workstation node. The sub-cloud tier costs about a third of a stand-alone cluster to own because one operator covers all sites.

8.3 Unit economics

At the §5.1 volume (500,000 pages per site per year, 2.5 M over 5 years):

Operations, not hardware, is the largest cost line in every tier except the workstation node.

Tier	5-year cost	Cost per page	Cost per 5-page case
Single-site cluster	$308,000	$0.12	$0.62
Sub-cloud, N = 4	$106,600	$0.04	$0.21

A 2-GPU node has headroom well beyond 2,000 pages a day for document models of this size, so cost per page falls almost linearly with volume until the GPUs saturate.

8.4 Sensitivities

- **Operator rate.** At an India platform-engineer rate of $35,000 per FTE-year the ops lines fall by 55%: the single-site cluster drops to about $196,000 and the sub-cloud to about $79,000.
- **Number of sites.** Ops per site falls with N. At 10 sub-clouds it is 0.05 FTE, $20,000 over 5 years, and the central node's share is $4,000.
- **Volume.** Hardware and ops are fixed until the GPUs saturate. Doubling pages halves cost per page; halving them doubles it. A site

below a few hundred pages a day belongs on a workstation node or a neighbour's sub-cloud.

- **Utilisation.** An idle GPU costs the same as a busy one. Consolidate low-volume sites before buying more hardware.
- **Hardware refresh.** A year-4 refresh adds the hardware line again: $95,000 to the cluster, $50,000 to the sub-cloud. If a workload needs an unquantised 70B model, an H100-class node replaces the L40S node at roughly $90,000–110,000 installed; the hardware line doubles and the finding on operations does not change.
- **The placeholders.** WRCP is priced per node, so it does not grow with GPU count or workload density. Integration and support depend on how much of §6 the site team takes on itself.

The inputs are exposed in a companion spreadsheet so each organisation can substitute its own volumes, hardware quotes and labour rates.

9. Getting started

Phase	Weeks	Activity	Exit criterion
1. Classify and size	1–3	Inventory use cases; classify data per §1.2; map workloads to the eight classes; size GPUs per site; run the §8 model on real volumes	Classification matrix; site bill of materials; TCO on own numbers
2. Develop centrally	3–8	Build and evaluate models on de-identified data on the central cloud (or public cloud for non-PHI); establish the signed-artefact promotion pipeline	Validated model artefact; evaluation report
3. Stand up the first sub-cloud	6–12	WRCP day 0 and day 1 via Conductor; deploy the clinical AI layer; dry-run on synthetic data; verify audit trail	Site attested; synthetic end-to-end run
4. Go live and template	12–16	Promote the model; enable HITL; run supervised; capture the blueprint for site N+1	First production site; replication blueprint

The second site is where edge programmes succeed or stall. The blueprint from Phase 4 is the deliverable that matters.

Appendix A: Models referenced in Section 5

Model class	Examples	Footprint	Used in
Vision-language, document understanding	MedGemma 1.5 4B; MedMO 8B	Single 24–48 GB GPU	§5.1
Mid-size clinical text generation	OpenBioLLM 70B; Meditron 70B; Llama 3.x 70B	2× 48 GB GPUs quantised (INT4/FP8); 2–4× 80 GB GPUs unquantised	§5.2
Entity detection for anonymisation	ClinicalBERT-class encoders; RadGraph	CPU or small GPU	§5.2
Small general-purpose with structured output	Gemma 4 E4B / 26B MoE; GPT-OSS 20B	Workstation GPU to single 48 GB card	§5.3
Embeddings for RAG	MedEmbed; bge-m3	Small GPU	§5.1, §5.3
Medical speech-to-text	MedASR	Single modest GPU	optional, §5.1

None of these is an approved medical device. Validate locally before use.

ABOUT PACE WISDOM SOLUTIONS

pacewisdom

Pace Wisdom Solutions is an IT services and software engineering firm of around 150 people, working across financial services, pharma and clinical data, and logistics. Its three practices are GenAI, Product Engineering and Cloud. In life sciences it builds and runs clinical AI workloads on customer-owned infrastructure: document intake, narrative generation and anonymisation, and SOP copilots at research sites.

pacewisdom.com

ABOUT WIND RIVER

WNDRVR

Wind River has built operating systems and platforms for mission-critical systems for four decades, from VxWorks and embedded Linux to Wind River Cloud Platform, a StarlingX-based private cloud that runs containers and virtual machines on commodity hardware from the far edge to the data centre. Conductor automates day 0, 1 and 2 operations across distributed sites; Analytics provides fleet telemetry.

windriver.com