THE CAPACITY ERA

Storage Days Tallinn

31 August – 1 September 2026

TALLINN, ESTONIA

The Dormant Corpus

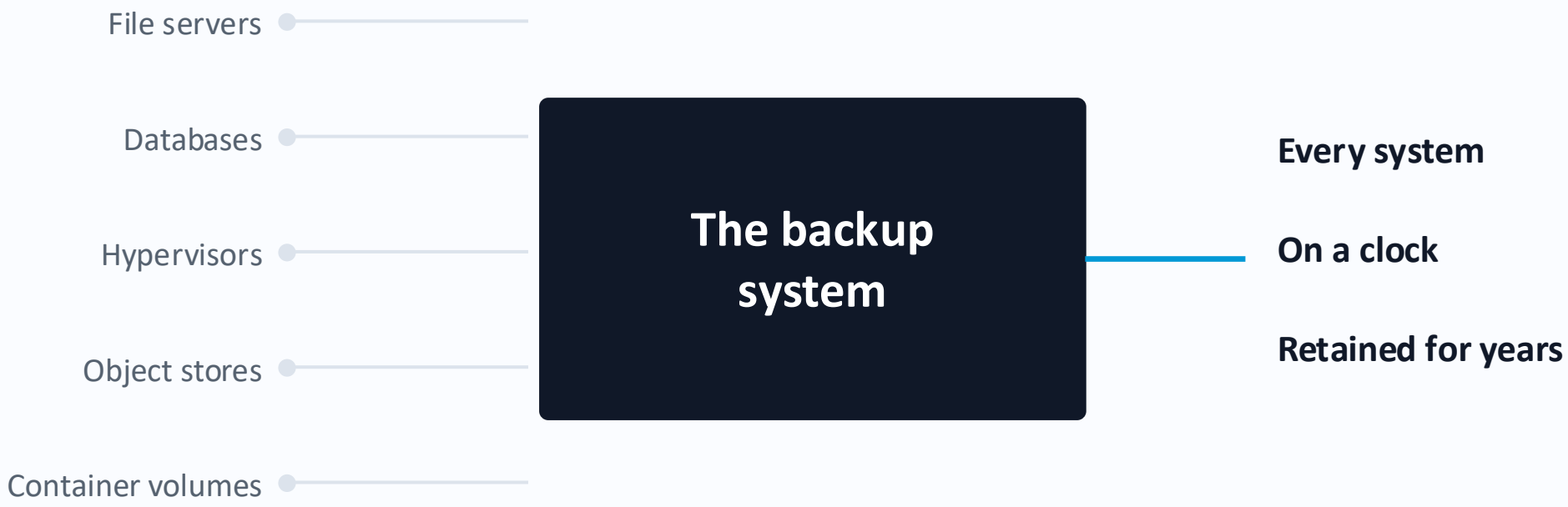
Running AI and analytics on backup storage

Rick Stevens SVP of Global Sales · Bacula Systems

leil

§1 OPENING

The one system that sees everything



Exactly one system touches every protected primary system, on a recurring schedule, and keeps the result with timestamps.

§1 OPENING

The only complete longitudinal dataset

Monitoring	Observability	SIEM	Data catalog
sees metrics	sees traces	sees logs	sees registered assets

The backup pass sees the actual bytes, of every protected system, on a clock, indefinitely

…and in nearly every data center, that corpus answers exactly one question: did something break, and can we put it back.

§2 GUARDRAILS

What I am not arguing

- ~~Analytic workloads on the active recovery chain~~
- ~~The production repository as a queryable data lake~~
- ~~Abandoning 3-2-1-0 survivability discipline~~

3-2-1-0

Three copies. Two media. One off-site.
Zero errors, verified continuously.

Necessary — and no longer sufficient.

The zero was the field's first step beyond pure survivability.

There are more steps.

§2 GUARDRAILS

A shift in framing

Longitudinal evidence layer not canonical record

DOES NOT CONTAIN

- Cache and ephemeral state
- In-flight transactions
- Log streams below the snapshot interval
- Anything retention has aged out

IS

- Bounded by what retention preserves
- Structured by what the schedule captured
- Complete enough, for long enough, to do what nothing else can

§3 WHY IT STAYED DORMANT

Three candidate answers. Two are wrong.

Compute ×

Cheap enough for a decade.

Capacity ×

The thing this room is actively solving.

Index

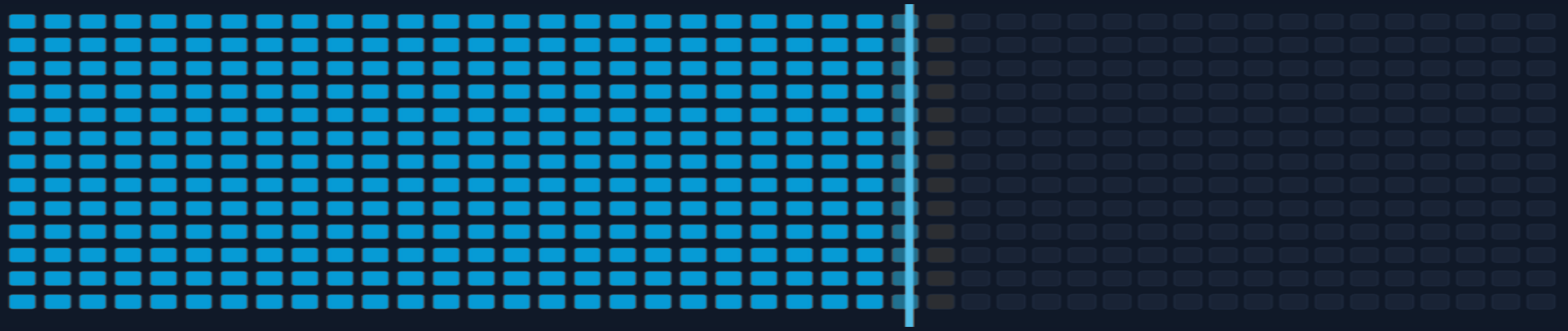
There isn’t one.

Nobody knows what is down there.

The backup catalog knows which job wrote which chunk to which volume. That is an operational ledger — not an index of content.

§3 WHY IT STAYED DORMANT

You cannot walk an exabyte



- Random sweep on media that wants sequential writes
- Wakes every drive the tier was keeping dark
- And again next month — the corpus keeps changing

The crawl is not a performance problem.

It is an architectural contradiction.

§3 WHY IT STAYED DORMANT

What a walker cannot reach

Already solved

Filesystem indexers walk metadata at scale, and read lake-format headers, footers, zone maps and range statistics in place. Extraction during a walk is a solved problem.

The consistency boundary

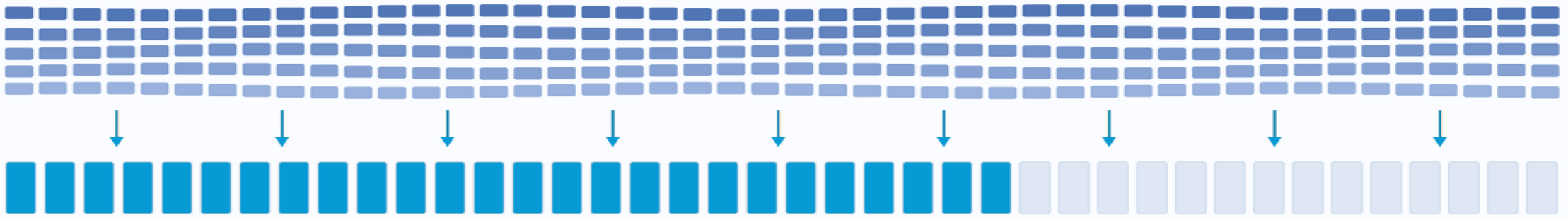
Databases mid-transaction. VM images. Object stores with no tree to walk. SaaS estates with no filesystem at all. A walker cannot open them. A protection pass already has to.

The time axis

A walk describes now. Run it again next month and you have two nows, not a history. Nothing in the walk records what happened in between.

A live index and a historical index are the same idea, one axis apart.

Index at ingest



READ · CHUNK · FINGERPRINT · VERIFY · STORE

The index falls out as a by-product — append-only, time-ordered, written into sequential zones

Stream shape = media shape.

Random-write index churn is a property of traversal, not of ingest.

§3 WHY IT STAYED DORMANT

The read is already paid for

A system that verifies its own backups already reads every byte it protects.

The index does not need a new pass over the data. It needs a tap on the one that already runs.

Every byte

no sampling, no exclusions

On a clock

the schedule already exists

Already computed

chunk boundaries, fingerprints

Already governed

under retention and key policy

§3 WHY IT STAYED DORMANT

Three axes of the index

Space

What exists, where, with what metadata.

owned by the Filesystem index

Time

What existed when. What changed between two points.

owned by the Protection pass

The unwalkable corpus

Databases behind a consistency boundary. VM images. Object stores. SaaS estates.

owned by the Protection pass

One index — not two vendor categories.

§4 AXIS ONE

Semantic — what do the bytes mean?

01	**Retrieval over history**	“What did engineering know about defect class X in Q2 2023?” stops being an archaeology project.
02	**Similarity indexing**	The signed contract and the drafts before it. A paper and its preprints. Datasets from a common source.
03	**Anomaly detection**	Hashing tells you a file changed. An embedding-space delta tells you whether it changed in a way that should worry you.

Open: false-positive rates against real workloads, embedding budget, model drift across the retention window.

§4 AXIS ONE

Structural at ingest. Semantic after.

AT INGEST

Structural index

- Cheap
- Stable
- Eligibility question is trivial
- Space, time, existence, integrity, classification hooks

POST-BACKUP, SELECTIVE

Semantic layer

- Expensive
- Volatile — models turn over quarterly
- Eligibility is the hardest question in this talk
- Immutable copies, decoupled schedule

Two artifacts. Two lifecycles. Conflating them is where most designs go wrong.

§4 AXIS ONE

Substrates for retrieval

Live tier

Current data is the point

- Customer service assistants
- Real-time fraud
- Operational queries

Backup tier

History is the point

- E-discovery
- Forensic reconstruction
- Regulatory inquiry
- Compliance attestation
- AI provenance

Complementary, not competitive. Where history is the point, the backup tier is the only honest substrate.

§5 AXIS TWO

Structural — what are the bytes?

Classification

PII, PHI, controlled and export-restricted material. A property of the index, not a project — and the precondition for every eligibility decision.

Lineage

Where a dataset came from, which jobs produced and consumed it. Open: reconstructing edges from incomplete metadata.

Attestation

Not “is this control in place” but “can you produce an immutable, time-indexed record that it was in place continuously.” That is a query.

You cannot enforce “never embed personal data of EU residents” if you do not know where it is.

§5 AXIS THREE

Behavioral — history as a time series

~2,500

temporal samples per object,
across tens of millions of objects

- **Configuration drift** — Below the layer infrastructure-as-code reaches.
- **Retrospective IOC scanning** — Scan history the moment a new indicator is published.
- **Integrity erosion** — The 2009 document opens. Does it render the same?

The most under-exploited large-scale dataset in production storage today.

§5 AXIS THREE

The diff nobody has run

01	**Configuration drift**	The delta between this month’s tree and last month’s — below the layer infrastructure-as-code reaches, and invisible to anything that only sees current state.
02	**Authorization moved**	A permission edge that changed without a ticket. No event, no log line, no owner — but a clear shape in a diff between two points in time.
03	**Mass mutation**	Re-encryption, bulk relabeling, staging before exfiltration. Each looks ordinary in a snapshot and obvious in a diff across the retention window.

Tree-diff is not a new capability. It has been run between two systems, or a system and its replica — almost never between a system and its own past, because the past was not indexed.

§6 AXIS FOUR

Scientific — the de facto active archive

1

The boundary is not where the diagrams say

Checkpoints are written to primary, captured by the protection pass, then accessed — sometimes exclusively — through the secondary tier.

2

Reproducibility lives in the catalog

The only system capturing data corpus, code repository and configuration at a single coherent point in time.

3

The audit substrate already exists

Regulators on three continents are converging on training-data provenance. The catalog already records where and when.

We are, more or less by accident, sitting on the audit substrate the AI governance community is trying to invent from scratch.

§6 AXIS FOUR

The selection economy

With an index, you read the highlighted fraction. Without one, you copy the whole corpus — and pay for it twice.

TODAY

- Promote everything to flash
- A second copy in a lake
- A third in a vector store
- A parallel ETL path to sync them

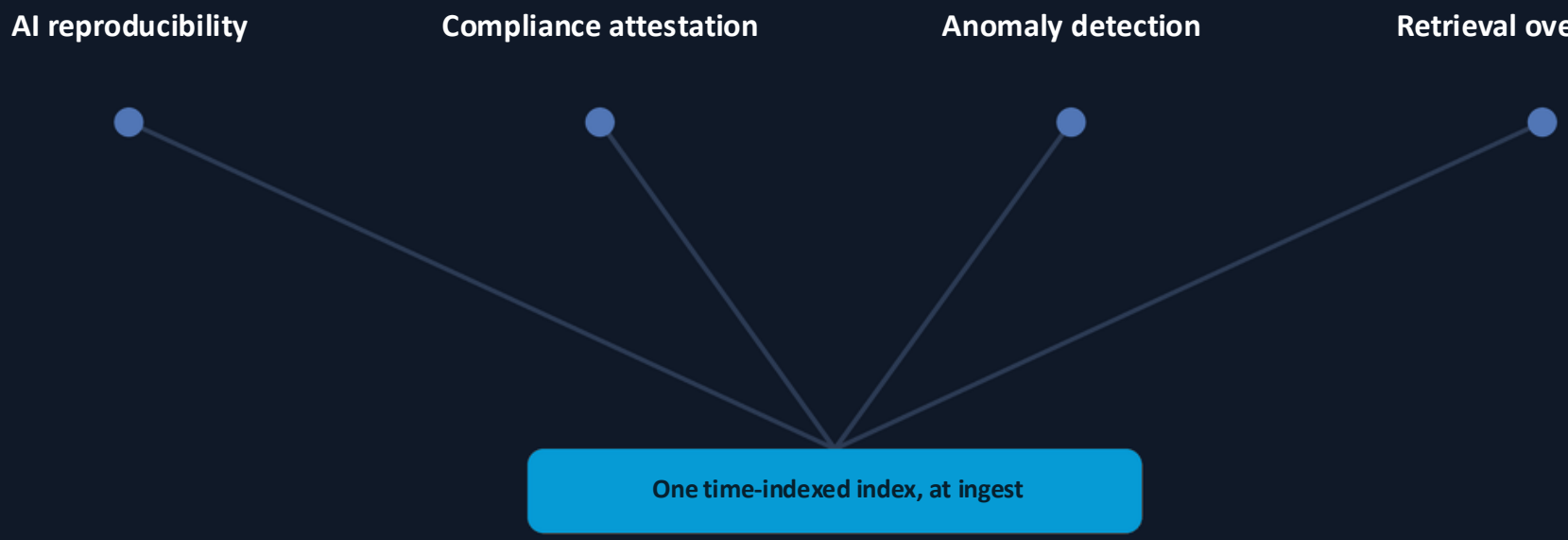
WITH AN INDEX

- Sequential read — what this media is good at
- No duplicate tier, power or carbon
- Read the fraction that matters, in place

The bottleneck was never bandwidth. It was selection.

§7 SYNTHESIS

One index. Four classes of query.



You do not have to solve this four times.

Sold today as four products, by four vendors, each building its own copy and its own index. An accident of how the industry grew up — not a property of the problem.

§7 SYNTHESIS

An emission, not a destination

A destination

A query interface bolted onto a backup product. Useful to one customer at a time, invisible to everything else in the estate.

the tempting design

A described feed

Record types, field definitions, a policy filter applied at the point of emission, and a documented mapping into the retrieval index the consumer already runs.

what the producer owes

Many consumers

Search, data catalog, compliance attestation, model-training provenance. Four consumers, one emission — not four products with four copies.

who it is for

If the index only works inside one product, the crawl gets built anyway.

Five constraints

Pricing

Capacity pricing penalizes computation.

1

Architecture

Compute near the data, selectively.

2

Storage substrate

Zones, power, deadlines.

3

Isolation

Never the recovery chain.

4

Sensitivity

Harder than the other four.

5

The last two are what distinguish a defensible reframing from a reckless one.

§8.2 ARCHITECTURE

The compute plugin pattern

THE ASSET

Active recovery chain — highly isolated, immutable, recovery-first

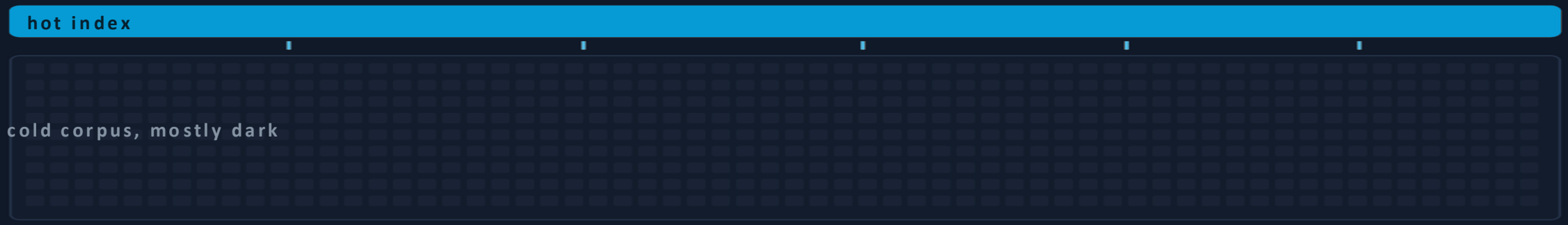
THE DERIVATIVE

Analytic pipeline — immutable copies, decoupled schedule, applied selectively

hard boundary

Open: plugin contracts, observability, failure semantics, isolation guarantees.

Constraint: the storage substrate



Zones

Host-managed sequential zones reward an append-only, time-ordered index — and punish random-write churn.

Power

Capacity powered down by default only works if the index that decides what to wake sits on a small, always-available footprint.

Deadlines

Bounded command latency makes a read schedulable rather than stochastic — but only when the request is small enough to bound.

What is the right ratio of hot index to cold corpus at exabyte scale?

If most queries resolve inside the index, the tier stays dark and the economics hold. A systems result, not a product feature.

Constraint: isolation

The recovery copy is the asset.

The analytic artifact is derivative.

The two never run in the same room.

No shared compute	No shared scheduling	No shared mutation rights	No shared failure domain

If the analytic layer shares any failure domain with the active chain, it has destroyed more value than any analysis could create.

§9 SENSITIVITY

The burden of history

Historical data is more sensitive than live data, not less.

The live tier reflects current access controls. The backup tier, precisely by doing its job correctly, preserves the prior state of all of them.

Revoked entitlements

as they once existed

Erased records

as they once existed

Declassified material

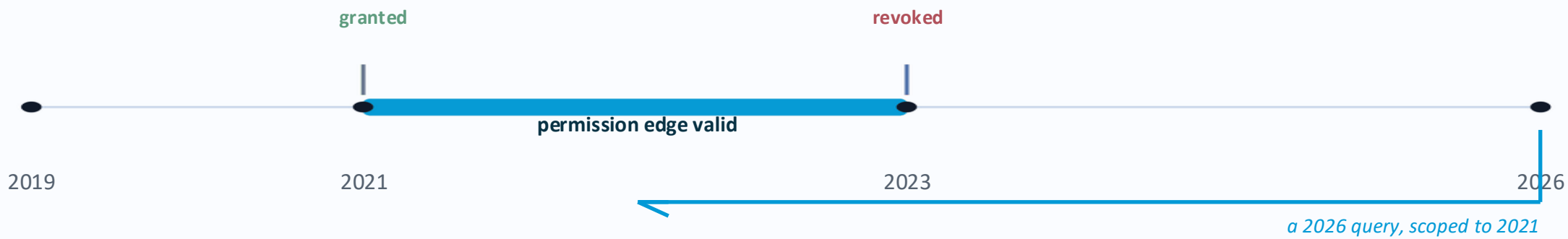
in its prior classified form

Personal data

of people who left years ago

§9 SENSITIVITY

Access control has a time axis



Querying forward

They had access in 2021 and do not today. Should a 2026 query return it? Almost certainly not — but the index was built when they did.

Querying backward

They have access today and did not in 2021. The investigator says yes. The DPO says it depends on lawful basis.

Erasure inheritance

A semantic index is a new, queryable artifact derived from data somebody asked you to delete — and it did not exist when they asked.

Most index structures have no way to express that a permission edge has a validity interval at all.

§9 SENSITIVITY

Two gaps nobody has closed

GAP ONE

Identity history is lost

- The data is retained; the state that governed it is not
- Who was in which group in 2021 is, in most estates, gone
- Nothing prevents retaining it
- Almost nobody does

GAP TWO

Wrong enforcement layer

- The filesystem enforces access — and knows exactly one present
- No query can weigh a 2021 credential against a 2026 one
- Workaround: historical identity state mounted beside a historical index
- A separate room per point in time. Not a query.

Worked once, before POSIX access control settled. Not revisited at this scale.

§9 SENSITIVITY

Selective exclusion and governance

Never computed, by policy —
not already computed, then suppressed.

Encryption boundaries

Derivatives inherit the sensitivity of their sources — and need their own keys.

Privilege separation

The restorer cannot query the index. The analyst cannot mutate the chain.

Long-horizon key management

Designed for the longest retention window in the corpus. Post-quantum ready.

§10 RESEARCH AGENDA

Six questions

1 Index-to-corpus ratio

What fraction of realistic queries resolve without waking the capacity?

2 Cost of indexing at ingest

What is produced inline, what is deferred, and how does that shift on zoned media?

3 Access control with a time axis

Permission edges with validity intervals, retained identity history, inherited erasure obligations.

4 Querying time-indexed metadata

What primitives, and what does an auditor accept as proof?

5 Anomaly detection in embedding space

False-positive rates against real production workloads.

6 Federation across boundaries

What “stays local” means when asserted about an index, not the data.

None of these are vendor questions.

3-2-1-0 is the floor, not the ceiling.

The index cannot be crawled into existence on the storage we are all now building.

It has to be a by-product of protection.

Backups remain recovery-first, regardless of any analytic processing we layer above them.

At Bacula Systems we are building this in two parts: an indexed protection pass (IDB) and a governed emission from the verify stream (BVI). Everything before this slide holds regardless of who builds it.

Storage Days Tallinn

31 August – 1 September 2026

TALLINN, ESTONIA

Thank you

Questions and discussion

Rick Stevens Bacula Systems

Rick.Stevens@baculasystems.com

leil