Is more cloud always the right answer?

>_ THE REPATRIATION CALCULUS

>_ THE 93% SIGNAL

93% of enterprises are actively repatriating AI workloads from public cloud.

Not evaluating.
Not piloting.
Actively doing it.

Source: February. 2026 Enterprise Infrastructure Survey

>_ THIS IS NOT CLOUD FAILURE

It is the economic phase change of mature infrastructure.

Phase 1 → Speed
Phase 2 → Scale
Phase 3 → Efficiency ← you are here

Repatriation happens at the Phase 2 → Phase 3 transition. Every time.

>_ 4 COST EVENTS THAT TRIGGER REPATRIATION

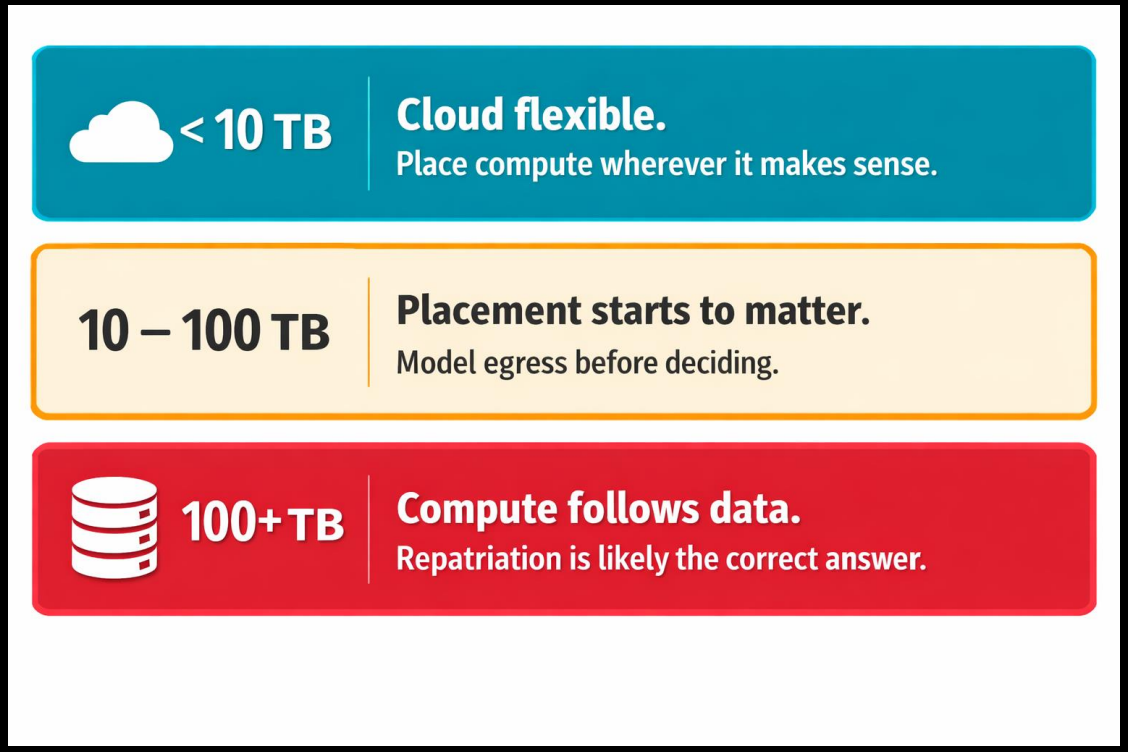
1. Egress Cost
   Per GB. Every time. No discount fixes the structure.

2. Licensing Overhead
   Cloud infrastructure premiums not in your license agreement.

3. Platform Tax
   Managed Postgres: $3k/mo. Self-hosted: $500/mo.

4. Operational Gravity
   The weight of every integration you'd have to lift to leave.

>_ DATA GRAVITY IS PHYSICS. NOT PREFERENCE.



Bottom line: At 100TB+, you don't move the data. You move the compute.

>_ WHY AI WORKLOADS REPATRIATE FIRST

GPU Utilization
Stable training pipelines favor owned GPUs.

Storage Throughput
I/O costs dominate at petabyte scale.

Iteration Frequency
Daily retraining cycles multiply every cost category.

Bottom line: AI teams repatriate first and fastest. The economics demand it.

⚠ THE 4 FALSE REPATRIATION FAILURE MODES

Ignoring staffing cost
On-prem needs engineers to run it. Model the headcount.

Underestimating hardware lifecycle
A $40k server costs $40k again in 5 years.

Skipping power and cooling 30-40% of on-prem AI infrastructure cost.
Always.

Rebuilding cloud complexity on-prem
The most expensive failure mode. Simplify. Don't replicate.

Bottom line: Repatriation succeeds when treated as an architecture project. Not a procurement exercise.

>_ THE FULL CALCULUS

Four cost events.

Data gravity thresholds.
Break-even modeling.
Decision framework.
False repatriation failure modes.

All of it — with comparison tables and a workload-by-workload decision framework.

Link in first comment ↓