阿里云
奥运会官方云服务合作伙伴

A New Dynamic PAMT Mode for TDX to Optimize the Metadata Memory Consumption for hyperscale deployment

Guorui Yu

ECS Confidential Computing Engineer, Alibaba Cloud

2026/03/12

Alibaba Cloud

Introduction

The Evolution of the Confidential Computing Tax

Striving for a "Confidential-by-Default" Cloud

The Goal

Tenants independently verify security/ privacy without relying on vendor trust.

The Reality

CC is currently a premium, opt-in feature, not the default.

The Barrier

The "Confidential Computing Tax".

The Heavy and Hidden Taxes of Previous TEEs

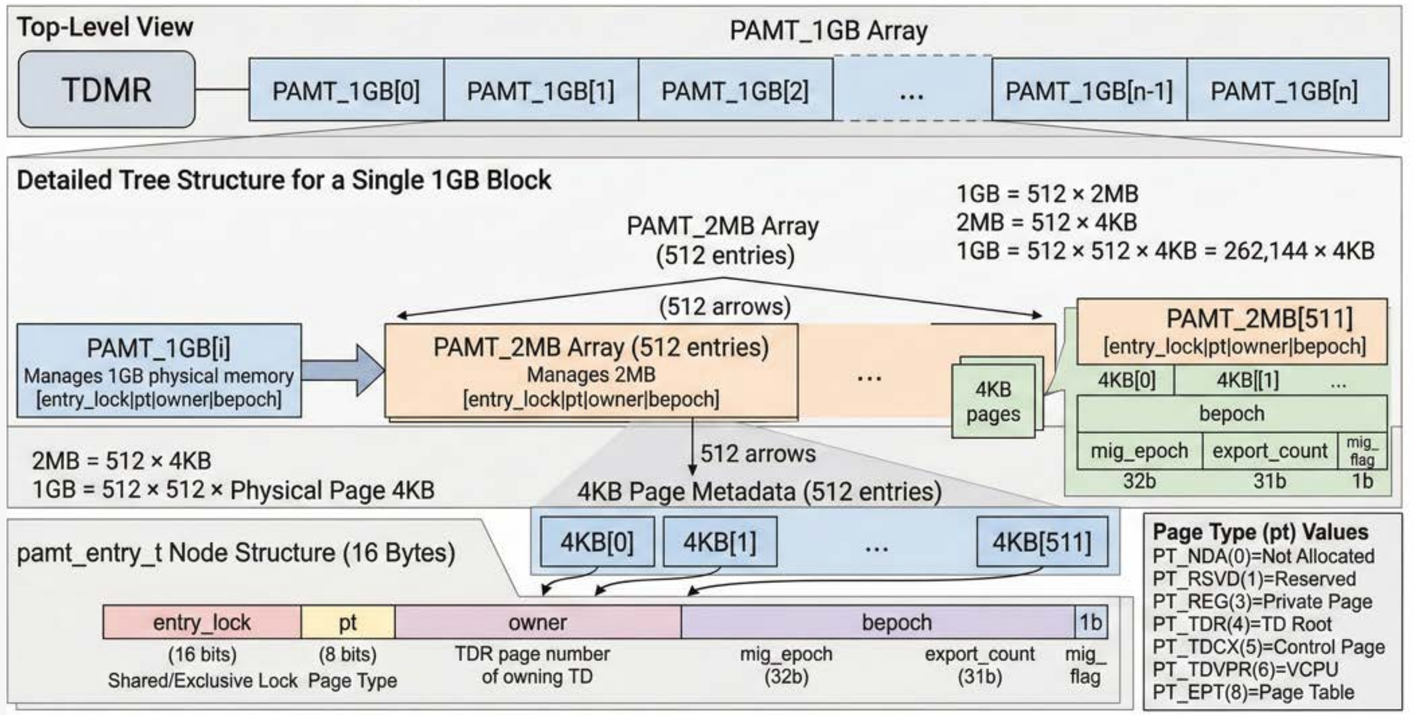
- Intel SGX (the rigid tax)
  - *The Leap*: IceLake expanded EPC (Enclave Page Cache) from MiB to GiB scale, a massive breakthrough.
  - *The Bottleneck*: EPC reallocation requires BIOS reconfiguration and a full system reboot. It cannot support rapid, dynamic resizing for cloud elasticity.
- Intel TDX & AMD SEV-SNP & ARM CCA (the static metadata tax)
  - Rely on strict hardware-enforced tables (PAMT for TDX, RMP for SNP, GPT for CCA).
  - Imposes a static metadata tax for every 4KB physical page, pre-allocated at boot.
  - You pay this tax even if you run **zero confidential VMs**.

The Hyperscale Dilemma & the Concrete Cost

- **The Illusion of "Small" Overhead**: A ~0.4% memory tax (4GiB per 1TiB) appears negligible on a single, isolated server.
- The Reality in Production (Alibaba Cloud ecs.g9i.48xlarge):
  - Spec: 192 Cores, 768 GiB RAM.
  - The Loss: Exactly 3 GiB of physical memory is permanently burned for PAMT at boot time, per server. Multiplied across a fleet of thousands of nodes, this results in Terabytes of unmonetizable, wasted memory for CSPs.
  - **Why does this matter for tenants?** It limits its accessibility for tenants — you may only have it in specific regions, specific availability zones, and with constraints on your elasticity.
- **The Target**
  - To make TDX “Default-on” a reality, eliminating this static memory tax was critical. Through Dynamic PAMT mode — which Alibaba Cloud is proud to have pioneered in production — we achieved a 98% reduction in PAMT memory consumption.
  - We thank the Intel engineering team for their dedicated support in making this possible.

Introducing the Dynamic PAMT Architecture

How Static PAMT Works (And Why It Doesn't Scale)



- Function: PAMT (Physical Address Metadata Table) tracks encryption and ownership metadata for every physical page assigned to a TD (Trust Domain).
- The Static Model: Historically, TDX enforces a flat, statically allocated array.
- The Constraint: A 16-byte metadata entry is rigidly required for every single 4KB physical page, leading to the 4GiB/TiB overhead.

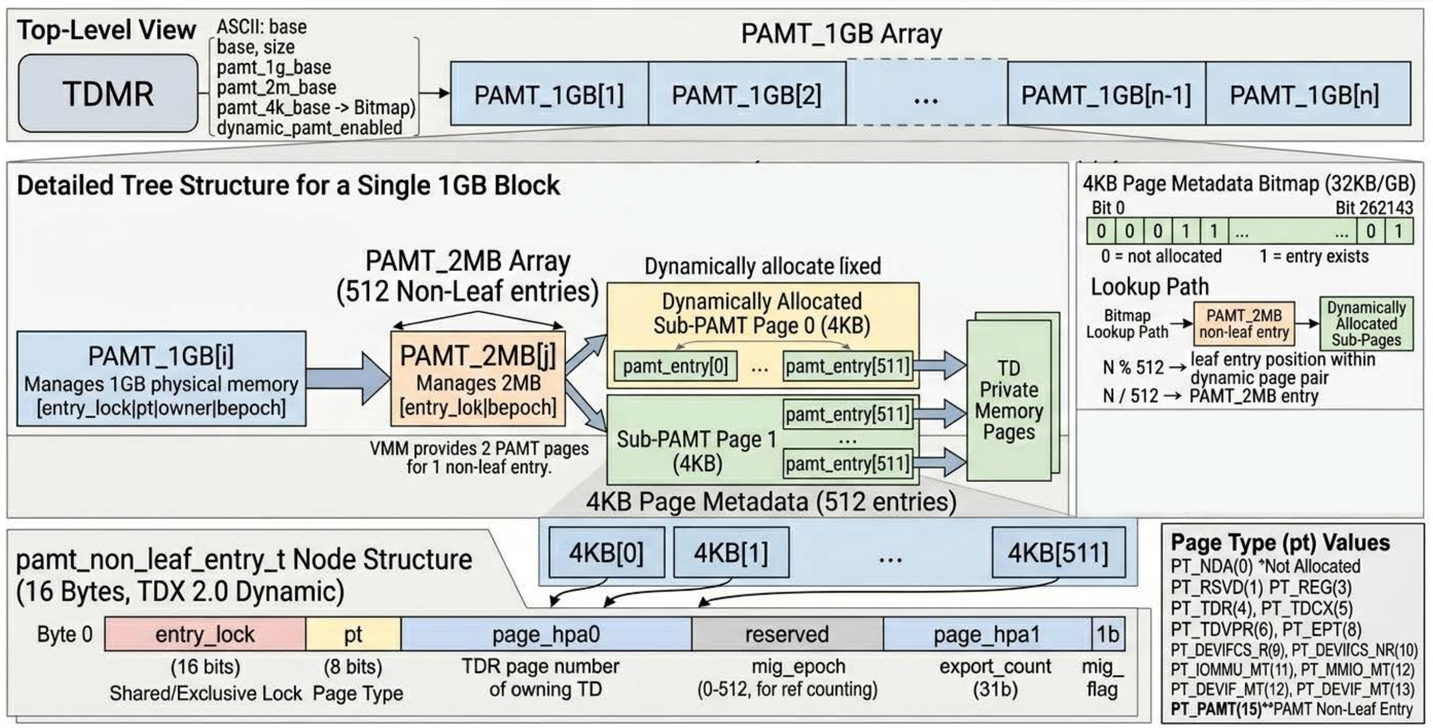
Aligning Security with Cloud Performance Best Practices

- **Tenant Indifference & The Hugepage Standard.**
  - Tenants care about security guarantees, not underlying mapping granularity. Furthermore, CSPs already default to providing hugepages (e.g., 2MB) to maximize TLB (Translation Lookaside Buffer) hit rates and boost overall system performance.
- **Extreme Sparsity & Stable Mappings.**
  - In real-world CVM, 4KB mappings are highly sparse. The vast majority of memory pages simply never require a conversion from 2MB down to 4KB during their entire lifecycle.

A Flexible, On-Demand Architecture

- Default to Large Granularity: PAMT entries are statically allocated at 2MB or 1GB levels by default.
- Eliminate Waste: This instantly cuts out the vast majority of meaningless 4KB-level metadata overhead.
- On-Demand Allocation: 4KB PAMT entries are only dynamically provisioned when strictly necessary.

The Dynamic PAMT Tree Structure



- **Tree Organization:** Dynamic PAMT operates as a logical tree structure, similar to page tables.
- **Static Upper Levels:** The PAMT_1G and PAMT_2M levels remain statically allocated.
- **Dynamic Leaf Nodes:** The PAMT_4K level is dynamically added and removed by the host VMM.
- **Page Pairs**: Tree nodes utilize pairs of 4KB pages to hold 512 PAMT entries for efficient management.

Triggers for Dynamic 4KB Allocation

- **TDX Control Structures:** Core management pages for the VM and vCPUs (e.g., TDR, TDCX, Secure EPT, TDVPR) inherently require 4KB metadata tracking.
- **Firmware Initialization:** TD VM boot sequence (e.g., TDVF init).
- **Memory Misalignment:** When guest memory mappings fall out of 2MB alignment boundaries.
- **Memory Conversion:** Dynamic transitions between private (encrypted) and shared (unencrypted) memory.
- **Peripheral Device Requirements:** Certain I/O peripherals or assigned devices may strictly require 4KB mapped pages for DMA or specific driver interactions.

Analyzing the New Overhead Model

- **The Static Base**: Extremely lightweight pre-allocated PAMT arrays for 1GB and 2MB levels, plus per-TDMR statically allocated Bitmaps.
- **The Dynamic Payloads**: 4KB Page Pairs, strictly allocated on-demand (e.g., during TDVF init or misalignment).
- **The Management Tax (KVM side)**: Host OS memory required to track and manage these Page Pairs (e.g., page tracking structures, metadata for the dynamic nodes).
- **Calculated Total**:
  - DPAMT_Mem $= \frac{P}{2^{26}} + \frac{P}{2^{17}} + \frac{P}{2^{15}} + \text{N2MiB} \times 2^{13}$ where $\left\lceil \frac{U}{2^{21}} \right\rceil \leq \text{N2MiB} \leq \min\left(\frac{U}{2^{12}}, \frac{P}{2^{21}}\right)$
  - P Total physical memory (TDMR size)
  - U Used TD private memory
  - N2MiB Number of 2MB regions containing at least one allocated 4KB page
- And the *Infinitesimals*,
  - SEAMRR Reservation: 32MB of memory permanently reserved for the TDX Module execution environment.
  - SGX EPC for Attestation: Remote attestation dependencies require enabling SGX in BIOS, forcing a 128MB EPC reservation per NUMA node (including subNUMA).
  - TME Hardware Latency: Hardware-level Total Memory Encryption inherently introduces a penalty of tens of nanoseconds in memory access latency.

Evaluation & Conclusion

Measuring the Impact on Alibaba Cloud

- Test Environment: Alibaba Cloud ecs.g9i.48xlarge (192 Cores, 768 GiB RAM), with 2MiB hugepage as the default config.
- Baseline Consumption: Static PAMT forced exactly 3 GiB of memory reservation.
- Dynamic PAMT Consumption: The consumption of PAMT page pairs is roughly in the range of 250-320 page pairs (2000KiB – 2560KiB).
- Net Savings: An empirical ~98% reduction in metadata memory overhead.
- Performance: +-0.5% Performance Implications on Sysbench/MLC

Delivering the "Confidential-by-Default" Cloud

- **Confidential Computing by Default** — Alibaba Cloud, in partnership with hardware vendors, is making secure-by-design cloud infrastructure the new standard.

- You can experience our TDX instances and perhaps the GPU confidential computing instances today in Singapore, Hong Kong, Beijing, and Hangzhou — with more regions coming soon.

Build a TDX confidential computing environment - Elastic Compute Service - Alibaba Cloud Documentation Center

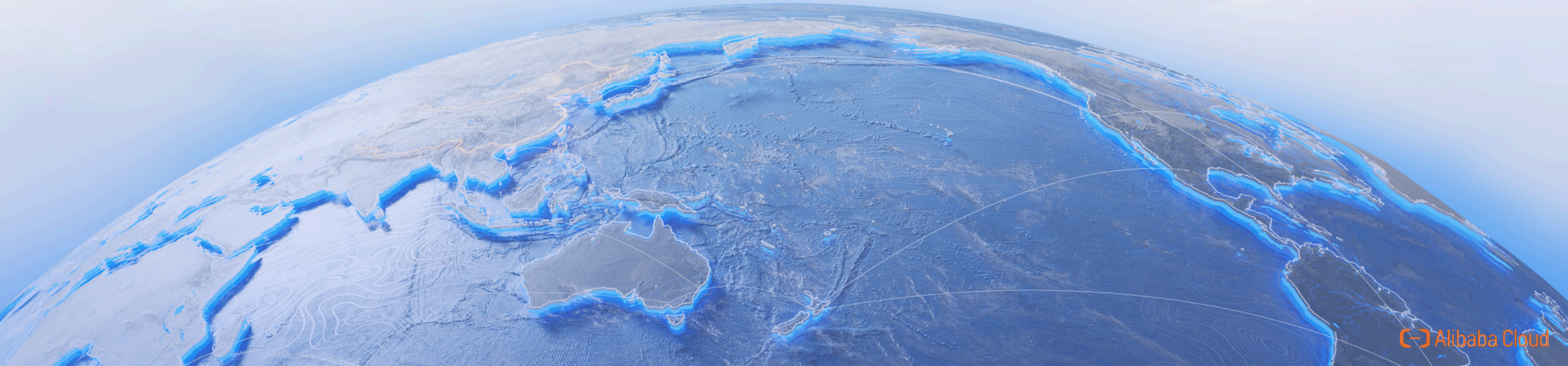
www.alibabacloud.com

Build a heterogeneous confidential computing environment - Elastic Compute Service - Alibaba Cloud Documentation Center

www.alibabacloud.com

阿里云
奥运会官方云服务合作伙伴

Thank you!



Alibaba Cloud