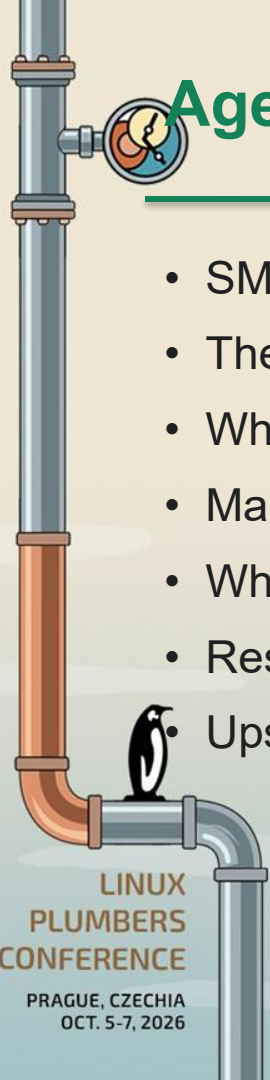


arm-smmu contig hint support and performance impact

Prakash Gupta · prakash.gupta@oss.qualcomm.com
Vijayanand Jitta Vijayanand.jitta@oss.qualcomm.com
Qualcomm India Private Ltd.

Linux Plumbers Conference 2026

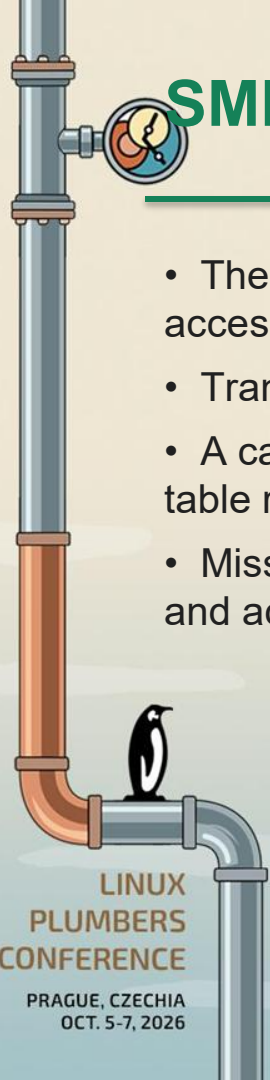


Agenda

- SMMU translation and TLB miss cost
- The contiguous (CONT) hint: what it is
- When CONT applies alignment and continuity
- Map / unmap behavior
- Why the SMMU PMU was needed
- Results: AI-inference workload
- Upstream status

LINUX
PLUMBERS
CONFERENCE

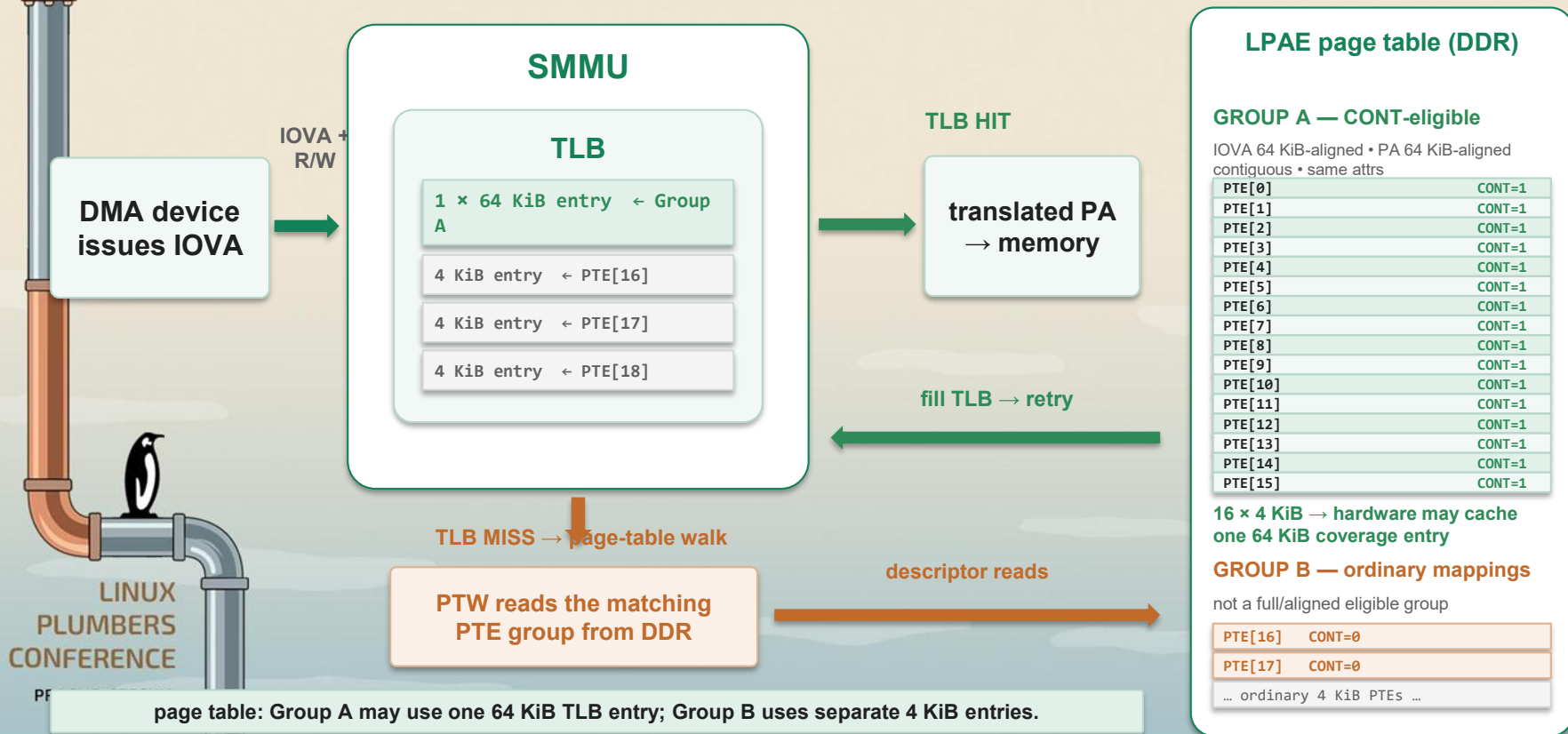
PRAGUE, CZECHIA
OCT. 5-7, 2026

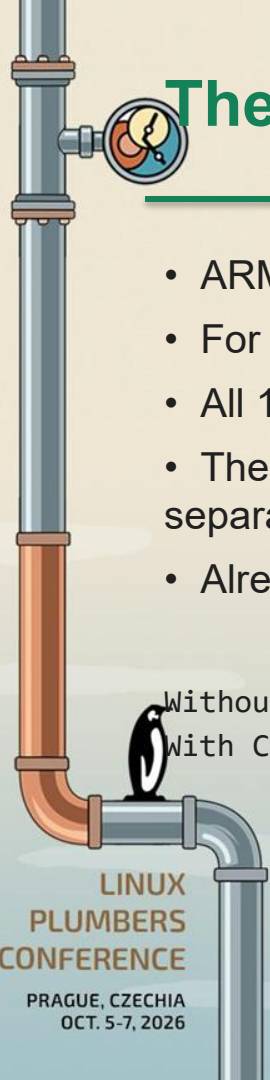


SMMU translation and the TLB miss cost

- The SMMU translates a device IOVA to a physical address before memory access.
- Translations are cached in the SMMU TLB / walk caches.
- A cache miss can trigger a page-table walk (PTW): extra table reads where page table reside in DDR
- Miss cost is implementation-defined; it depends on cache residency, table depth and access locality.

DMA access flow: CONT vs. ordinary mappings



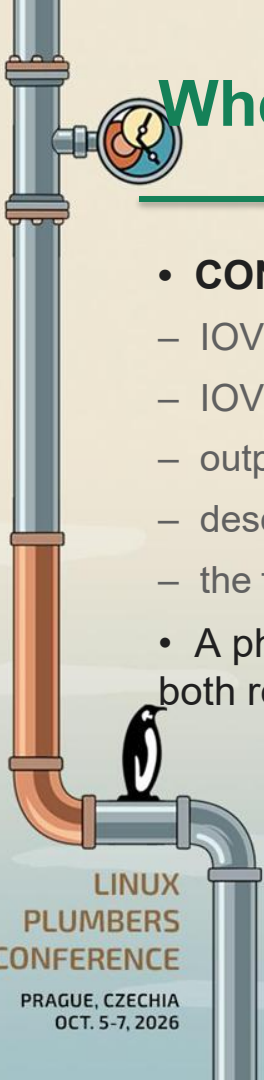


The contiguous (CONT) hint

- ARM LPAE descriptors carry a contiguous (CONT) bit.
- For a 4 KiB granule, the common group is $16 \times 4 \text{ KiB} = 64 \text{ KiB}$.
- All 16 descriptors remain in the table; each is tagged CONT.
- The hardware may then give the group combined TLB coverage instead of 16 separate entries.
- Already used on the CPU side; this work brings it to arm-smmu driver.

Without CONT : 16 independent 4 KiB translations cached

With CONT : 16 descriptors tagged -> combined 64 KiB reach

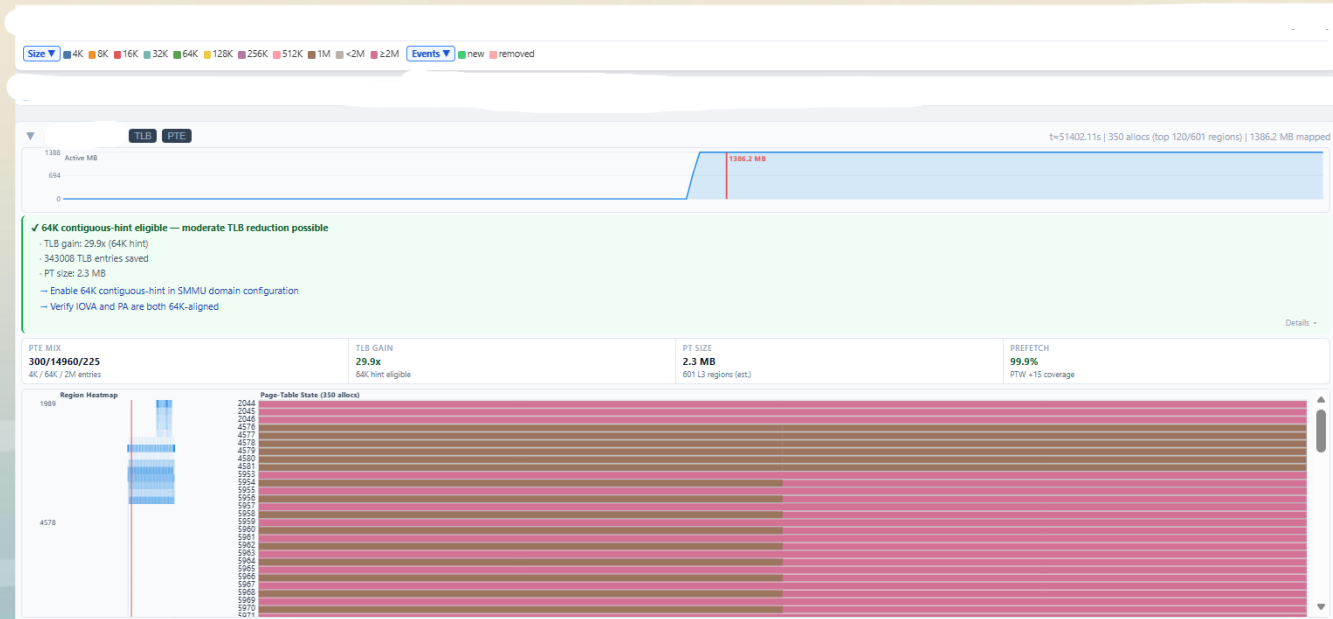


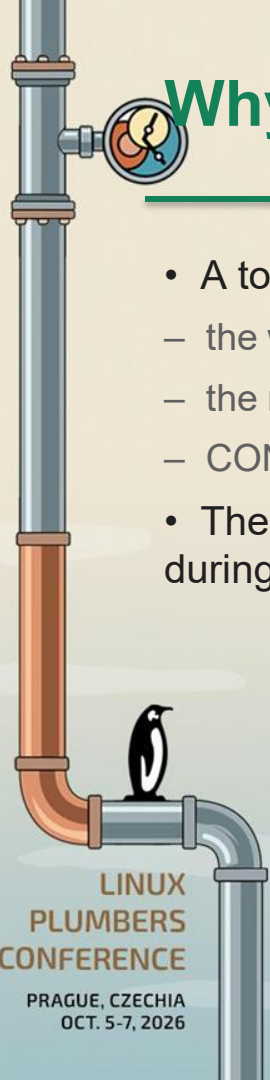
When CONT applies

- **CONT is set only when the whole group qualifies:**
 - IOVA aligned to the group size
 - IOVA range is contiguous
 - output physical pages are contiguous
 - descriptor attributes match
 - the full group is present
- A physically contiguous buffer alone is not sufficient; alignment and continuity are both required.

Map / unmap behaviour

- On map, an **eligible aligned** group is tagged CONT on every member.
- On unmap, non-partial unmap, of a tagged CONT group is cleared.





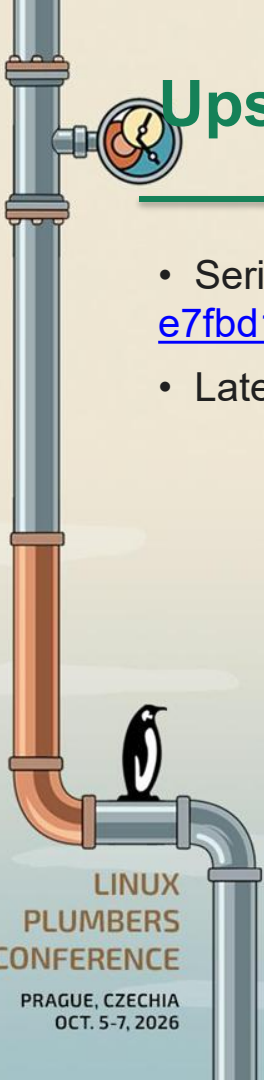
Why the SMMU PMU was needed

- A top-level on/off KPI result is ambiguous. A flat result could mean:
 - the workload had little translation pressure, or
 - the mappings were not CONT-eligible, or
 - CONT was not used effectively by the implementation.
- The SMMU PMU exposes lookup, hit and page-table-walk counters via perf during the run.

Results: AI-inference workload

Metric	Change with Contig hint
Throughput (Token per Seconds)	> +60 %
Page-table walks	↓
MTLB lookups	↓

- CONT-on runs show 64 KiB mappings and 92.8% leaf coverage
- Fewer TLB and walk-cache lookups, fewer page-table walks.
- Higher throughput on the same workload.

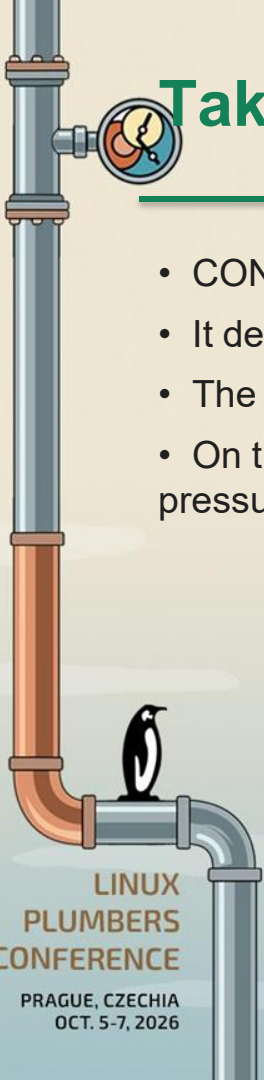


Upstream status

- Series: https://lore.kernel.org/all/20260921-iommu_contig_hint-v5-1-e7fbd1c3774d@oss.qualcomm.com/
- Latest revision: v5 (posted 2026-09-21), under review on linux-iommu.

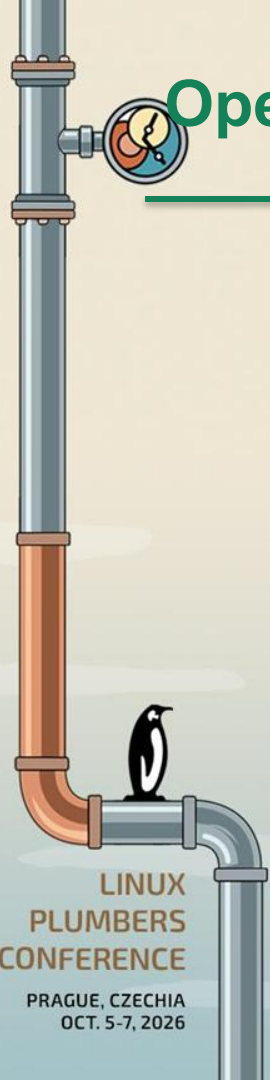
LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026



Takeaways

- CONT improves TLB coverage of eligible mappings.
- It depends on alignment, continuity and matching attributes; it is a hint, not a guarantee.
- The SMMU PMU turns an on/off experiment into an attributable analysis.
- On the measured AI workload: ~60% higher throughput with lower lookup and walk pressure.



Open questions

Current status

- Actively being reviewed in mailing list.
- No blockers for SMMUv2.
- SMMUv3 Opens: invalidation errata (MMU-700 #3777127 / SMMUv3 #3673557) and hardware dirty tracking (ARM_LPAE_PTE_DBM). [1]

[1] <https://lore.kernel.org/all/20260924225534.GD16465@ziepe.ca/>

Backup



LINUX
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026



Implementation details:

1 Advertise a usable size

`arm_lpaee_restrict_pgsizes()`

- starts with format-supported base sizes
- adds CONT group sizes to `pgsize_bitmap`

• **Opt-out:**
`IO_PGTABLE_QUIRK_ARM_NO_CONT_HINT`

2 Map an exact group

`__arm_lpaee_map()`

- core selects the exact group size
 - 4K: $16 \times 4K = 64K$; 2M: $16 \times 2M = 32M$
 - IOVA + PA aligned; PFNs contiguous; attrs match
- **set CONT (bit 52) on every base descriptor**

3 Unmap as a group

`__arm_lpaee_unmap()`

- inspect actual CONT at group edges
- partial group → reject before any PTE clear
 - rejected request is a complete no-op
- full group → clear/invalidate with TLB gather + BBM

16 descriptors remain in the table: CONT is a hardware coalescing hint, not page-table compression or a guaranteed single TLB entry.