



Item 6.2: Cloud-Based GPU Solution to Improve Congestion Revenue Rights (CRR) Application Performance – Implementation Update

Alfredo Moreno
Manager, Forward Markets

Sreenivas Badri
Sr. Director, Grid and Market Solutions (GMS)

Technology and Security Committee Meeting

September 14, 2026

Purpose

Implementation update on Cloud-Based GPU solution to improve CRR application performance.

For information only

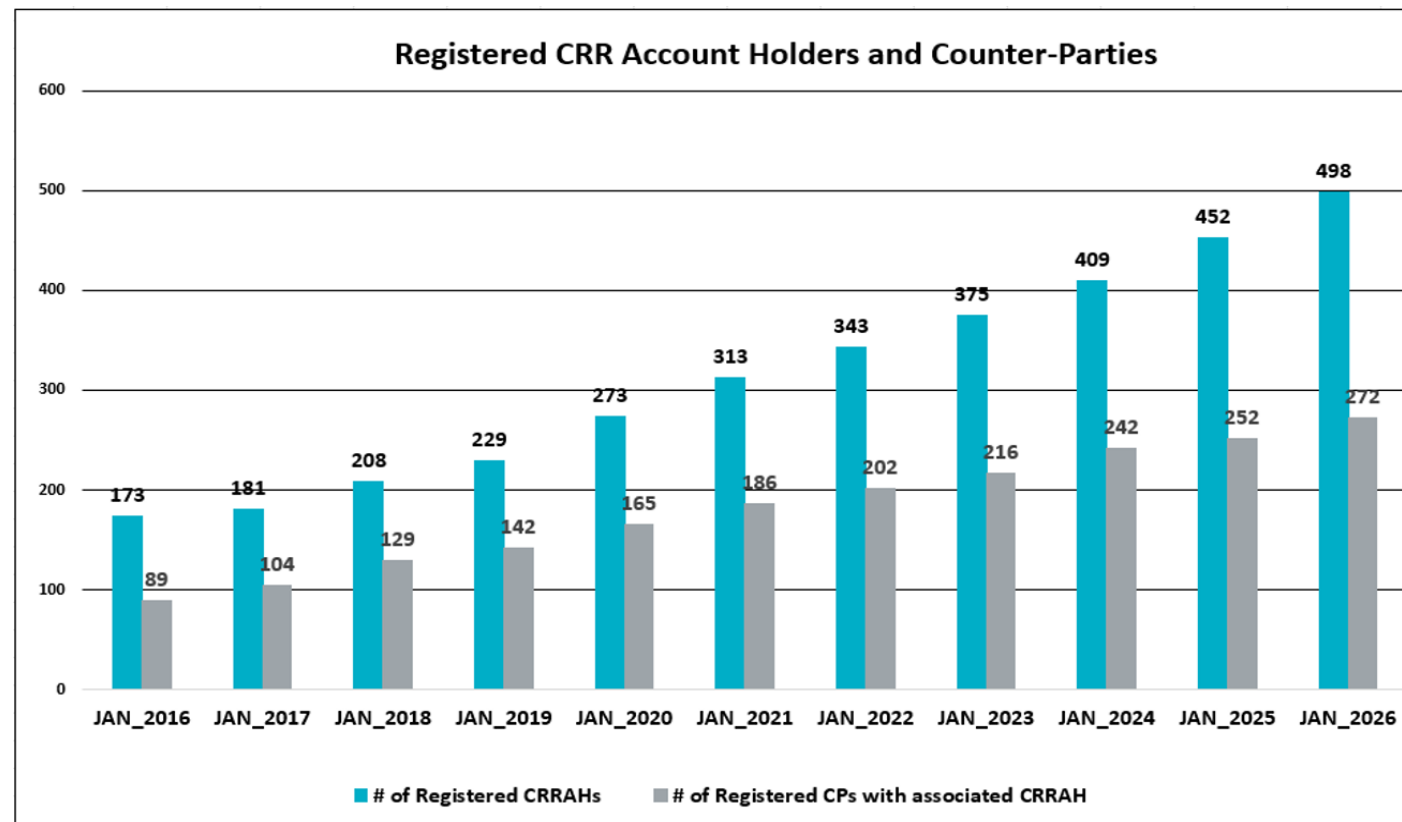
No action is requested; for discussion only.

Key Takeaways

- CRR auctions ran 5-10 days since Jan 2024 vs. a 2–3-day target; software fixes and current hardware tuning closed only part of the gap
- Cloud POC with GPUs cut solve times significantly – 2.4X to 3.25X faster
- Implemented Cloud-Based GPU Solution in ERCOT Public Cloud, enabling August CRR Long term Auction to complete successfully under 3 days.

CRR Participation Has Repeatedly Outpaced Prior Expectations

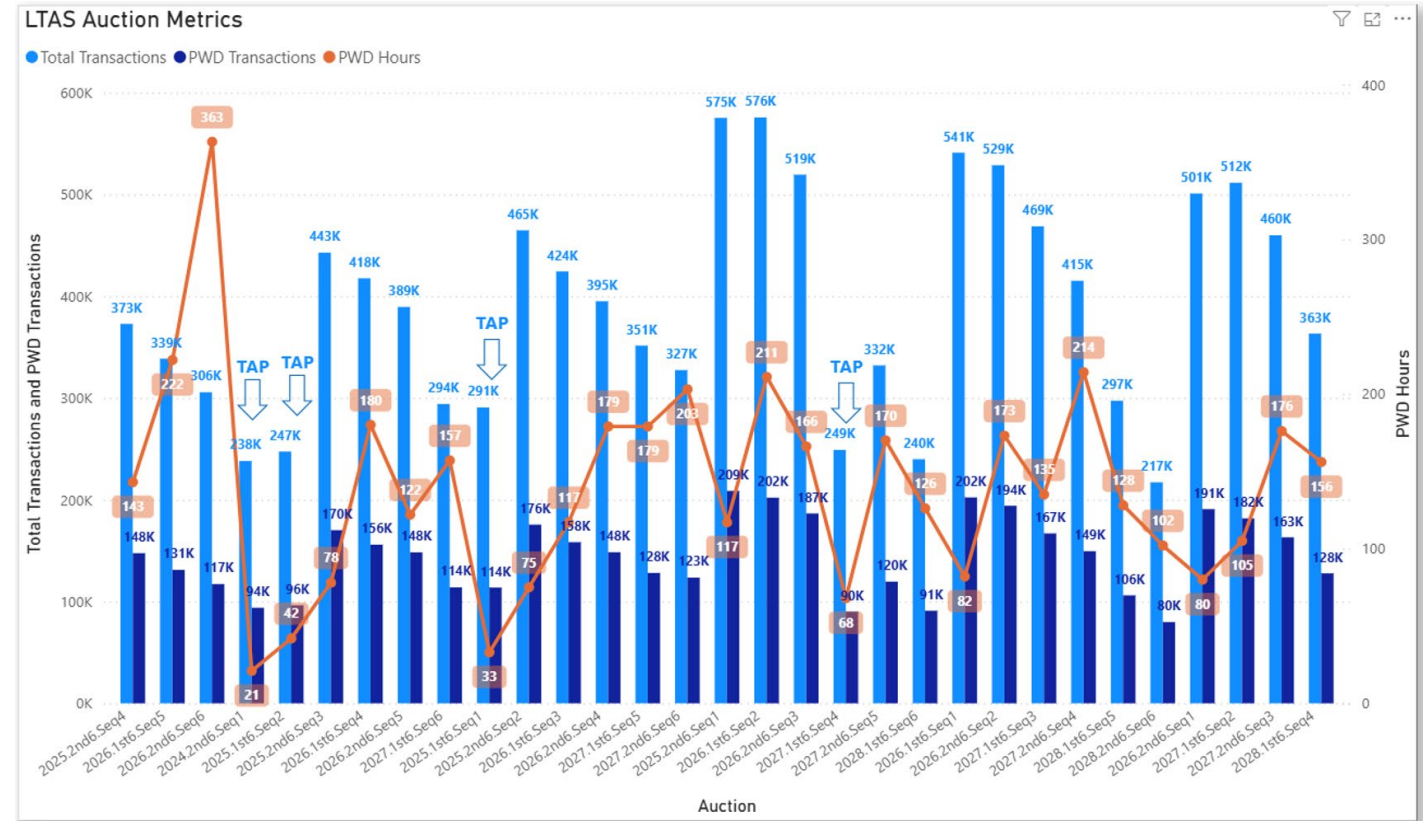
- **Oct 2019** – 2020.1st6.Seq1 becomes the first sequence to exceed 300K transactions, triggering the first Transaction Adjustment Period (TAP).
- **Dec 2019** – NPRR972 implemented, giving ERCOT discretion to declare TAPs.
- **Apr 2023** – 2023.2nd6.Seq1 is the first sequence to exceed 400K transactions.
- **Oct 2024** – 2025.1st6.Seq1 is the first sequence to exceed 500K transactions.
- **Apr 2025** – NPRR1261 implemented, allowing ERCOT flexibility in setting transaction limits. ERCOT has applied this flexibility to ensure transaction volumes remain within the CRR system's operational capacity.



Key Takeaway: Registered CRR account holders and counter-parties have nearly tripled since 2016, driving transaction volumes past the 300K, 400K, and 500K marks and requiring new rules (NPRR972, NPRR1261) to reduce TAPs and keep transactions manageable.

LTAS Auction Durations Kept Climbing Despite System Upgrades

- **Jan 2024** – 2025.2nd6.Seq4 is the first Long-Term Auction Sequence (LTAS) to exceed 100 auction hours.
- **Feb 2024** – 2026.1st6.Seq5 is the first LTAS to exceed 200 auction hours.
- **Mar 2024** – 2026.2nd6.Seq6 is the first LTAS to exceed 300 auction hours. First time since CRR go-live to publish late.
- **Apr 2024 – Dec 2025** – Hardware tuned, CPUs increased, and six vendor software releases delivered (Jun '24, Sep '24, Mar '25, Jun '25, Oct '25, Dec '25).
- **Solutions discussed** – Removing multi-month LTAS transactions, option pricing disclosure, running LTAS in parallel months, database cleanup, and cloud-based solutions.



Key Takeaway: LTAS auction hours repeatedly set new highs (100/200/300+ hours) as transactions grew, prompting ERCOT to evaluate multiple changes to avoid risks of late publication or potential failures. As our market evolves, increased transactions and market behavior will continue to impact runtime. Even after hardware tuning and six vendor software releases since 2024, runtimes remained high.

Root Cause Analysis

Problem Statement

- Since January 2024, CRR long-term auctions (LTAS) have taken 5-10 days to complete versus an expected 2-3 days
- The March 2024 LTAS took 363 hours, resulting in a missed auction-market posting deadline under the CRR Activity Calendar. The T&S Committee was updated at its April 2024 meeting.
- Several subsequent auctions came close to violating CRR auction posting deadlines.
- CRR business invoked the Transaction Adjustment Period (TAP) to cut bid volume and protect market timelines

Root Cause (18-month investigation)

- Organic network model growth (more nodes, more bid paths) is a contributing factor but not the primary driver
- Rising CRR Multi-Month Bid volume was confirmed by the vendor as the major contributor to the performance decline

363 hrs

Longest CRR LTAS run (March 2024) vs. 2-3 day
Target

18 mo.

To isolate root cause: network growth + rising
CRR Multi-Month Bid volume outpacing engine
capacity

Key Takeaways: CRR long-term auction performance has significantly deteriorated since early 2024. In addition to increased CRR transactions, the primary root cause is identified as the surge in CRR Multi-Month Bid volume, which has outpaced the existing engine's processing capacity

Solutions Explored: Software & Hardware

SW**Software Improvements**

- 18 months of vendor engine software changes, including algorithm improvements
- Meaningful gains delivered, but auctions still take 5-10 days as bid size and model complexity grow
- Vendor has exhausted further software and algorithm optimization options

HW**Hardware Improvements (On-Prem)**

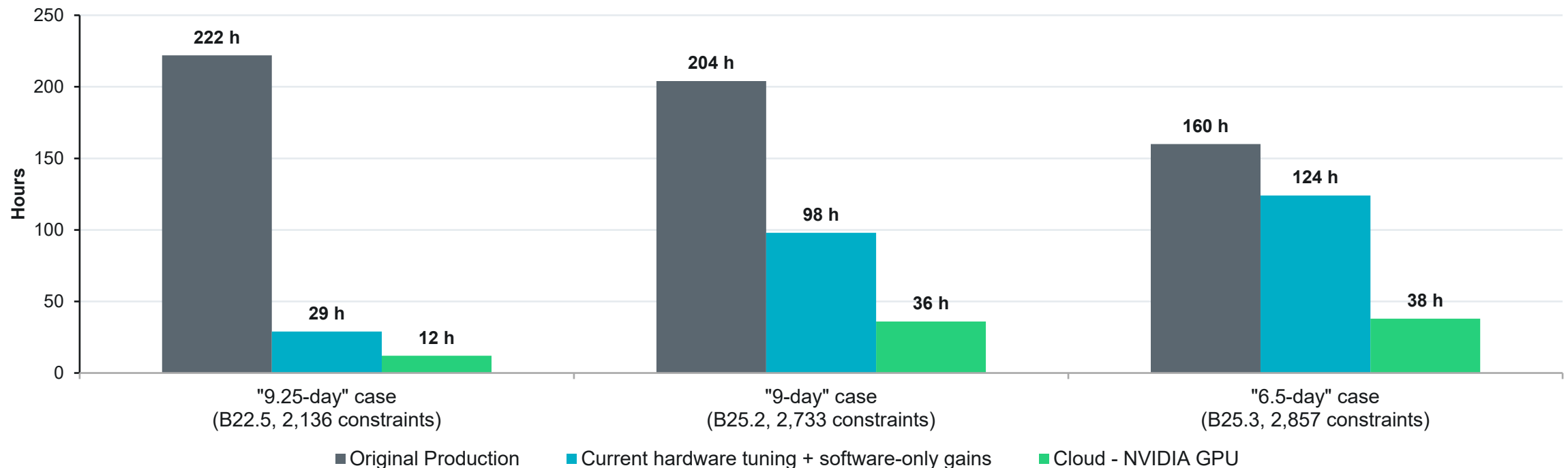
- April 2024: CRR engine server CPUs increased from 16 to 28, improving performance ~20%
- Adding more CPUs/VMs doesn't help further – CPU/RAM bus saturates on current hardware, and auction math is inherently sequential
- Advanced hardware configurations identified as the more promising path forward

Key Takeaway: Software and on-prem hardware upgrades delivered partial relief (~20-40%) but couldn't close the gap to the 2–3-day target – prompting evaluation of cloud-based GPU hardware

Cloud POC: Hardware Evaluated & Performance Results

- Vendor set up a POC environment on their public cloud to run CRR engines on Cloud with GPU technology.
- Vendor ran CRR Long Term Auction cases on public Cloud evaluating AMD EPYC, ARM, Intel Xeon, and NVIDIA GPU configurations, benchmarked against current hardware tuning + software-only gains and production.
- NVIDIA GPUs consistently delivered the best results across all long-term auction cases tested

CRR Long-Term Auction Solve Time by Hardware Configuration (hours)



Key Takeaways – Based on vendor Cloud POC results NVIDIA GPU solution was selected to run CRR engine in ERCOT's Public cloud.

Production Go-Live: CRR Auctions Now Running in the Cloud

1 Cloud Infrastructure Setup

- Partnered with Hyperscaler to procure NVIDIA GPU-powered hardware on ERCOT public cloud and establish Dev/Test/Production subscriptions

2 Validation Testing

- Tested multiple CRR long-term auction cases from on-prem production in the Cloud Dev/Test subscriptions – observed performance gains consistent with the Cloud POC

3 Production Go-Live

- Went live running CRR Long-Term Auctions in the ERCOT Public Cloud in August 2026

57 hrs

First Production CRR Long Term Auction (2028.2nd6.AnnualAuction.Seq5) started on 08/20 and completed in 57 hours in ERCOT Public Cloud

<3 days

First CRR long-term auction to finish under 3 days in the last 3 years

Key Takeaways: CRR Long Term Auctions are running in CRR engine servers hosted on ERCOT Public Cloud from August 2026, resulting in significant performance improvements. CRR UI and CRR Databases continue to operate on-prem.