RNP-021: Expansion of RNP-019 to Include Enterprise-Grade Compute

Status: Draft Submission
Date: November 04, 2025
Proposer: The Render Network Foundation
Category: Core Proposal – Technical Subcategory

Abstract

While the Render Network remains primarily focused on artists and their global network of consumer GPUs, some of the current leading image and video models cannot yet be run on consumer nodes. We have therefore expanded the scope of RNP-019 by introducing RNP-021, which will bring enterprise-grade nodes which can run these models in order to unlock access to and power these leading models, building directly on its framework for AI and general compute tasks. RNP-019 introduced trialing consumer-grade GPUs (RTX 4090, RTX 5090) with dedicated rewards for compute node operators powering AI and general compute workloads. RNP-021 extends this to include enterprise-grade GPUs (NVIDIA H100, H200, A100, L40, L4, T4, AMD Instinct MI300 series, Intel Data Center GPU Max series, and others if/when they become relevant, for example Groq LPUs), supporting aggregate compute capacity equivalent to up to 1,200 NVIDIA H200 GPUs through a flexible mix of approved hardware.

A unified compute emissions pool ensures equitable rewards, with adjustments for higher-performance hardware. This expansion enables the network to handle more demanding workloads, such as large-scale video and image generation models. This RNP initially maintains RNP-019's core mechanics for job allocation, rewards, and governance with two notable changes:

- Firstly, it proposes increasing the baseline job rewards for an RTX 4090 from 10 RENDER to 25 RENDER per epoch for the near term, and
- Secondly, it introduces a path towards a dynamic auction reward system that allows compute nodes to set their own asking rewards within the existing BME systems. This model creates a path for scaling the network.

Motivation

The global data center GPU market is projected to grow from USD 83 billion in 2025 to USD 353 billion by 2030, with a CAGR of 33.65%, driven by demand for AI, machine learning (“ML”), and High-Performance Computing (“HPC”) across industries like healthcare, finance, and media. https://www.mordorintelligence.com/industry-reports/graphics-processing-unit-market
Building on RNP-019's foundation, which addressed the need for scalable compute for the tools of today, RNP-021 responds to community and user feedback by incorporating enterprise-grade hardware. On the demand side, customers require higher-grade compute for diverse job types, including:

- **AI Training**: Large-scale training of foundation models, such as LLMs, image models, or world models with large dataset and VRAM requirements.
- **AI Inference**: Real-time tasks like chatbots, image generation, video generation, and more.
- **Video Generation Models**: Training and inference for models like ByteDance’s Seedream, Wan 3.1, and others, requiring high VRAM (80–141GB HBM3) and parallel compute for high-resolution, frame-sequential processing.
- **Image Generation Models**: Processing diffusion models (e.g.,Flux, Seedream) for high-fidelity image synthesis, demanding robust memory bandwidth.

RNP-021 expands RNP-019's vision by enabling the network to support these computationally intensive workloads—particularly for video and image generation models requiring 10–100x more compute than traditional LLMs. By building out this capability for the network, market and customer research can be gathered for wider use cases and provide further insights for adjustments moving forward. We anticipate that several existing and new partners will tap into this compute. One of the most active ecosystem partners of the Render Network, OTOY, additionally will commit to the usage of this compute cluster for their upcoming OTOY.ai platform for effortless AI image and video generation.

Specification

RNP-021 amends RNP-019's compute node requirements to include enterprise-grade GPUs in two distinct groups:

- Initially up to 100 enterprise nodes onboarded in the same way as RNP-019
- The balance of GPUs to be accessed as needed to match specific blocks of demand.

For the rollout and test phase of the alpha product, combined aggregate compute power equivalent to up to 1,200 NVIDIA H200 GPUs (~2,003 TFLOPS FP16 Tensor each in base mode, total ~2,403,600 TFLOPS FP16 Tensor for AI workloads; or ~67 TFLOPS FP64 each for HPC), which can be composed of a mix of approved hardware. Equivalence is benchmarked primarily on FP16 Tensor TFLOPS for AI/video models (prevalent in Render jobs), with FP64 as secondary for scientific

simulations. Compute power equivalency threshold can be revisited as demand and market dynamics change. Examples of mix for FP16 Tensor equivalence:

- **H100**: 1,200 H100s (989 TFLOPS FP16 Tensor each) = ~1,186,800 TFLOPS.
- **A100 (80GB)**: 2,000 A100s (312 TFLOPS FP16 Tensor each) = ~624,000 TFLOPS (scaled adjustment for full equivalence via quantity).
- **MI300X**: 1,500 MI300X (1,300 TFLOPS FP16 estimated each) = ~1,950,000 TFLOPS.
- **L40/L4/T4**: Significantly higher quantities (e.g., ~10,000 L40s at ~184 TFLOPS FP16 Tensor each) due to lower per-GPU performance.

The Render Network Team will verify cohort compute power using standardized benchmarks.
Note: Apple hardware (e.g., M5, M4 Max, M3 Max, M2 Max) is not included at this time.

Flexible Hardware Model Inclusion

To accommodate new hardware models, RNP-021 introduces a dynamic onboarding process extending RNP-019:

- **Performance Benchmarking**: Evaluate new chips based on metrics like TFLOPS, TOPS, and memory bandwidth benchmarks.
- **Customer Pricing Methodology**: Initially add the ability to set rates as necessary by the Render Network Foundation as for RNP-019 benchmarked to industry standards, while the project works towards a dynamic pricing system driven by node asking price for rewards and aligning with fluctuating market rates. For example comparing performance to H200 (baseline: ~100 TFLOPS FP64) and referencing competitor pricing from Runpod, Vast.ai, and Google Cloud where, for instance, a new cluster with 80% of H200's TFLOPS would tend to be priced at ~80% of H200's rate unless there were other material market factors in play.
- **Approval Process**: The Render Network's technical team reviews new chips for compatibility, with community input via Discord. Approved models are added to the pricing oracle as applicable.
- **Implementation**: Integrate new chips into the job allocation system, ensuring support for diverse workloads. For example, Groq LPUs (750 TOPS, 10x faster inference) could be added once decentralized integration is validated. https://groq.com/technology

Node Operator Requirements

- **Hardware**: Approved enterprise GPUs, revised as appropriate from time to time.
- **Bandwidth**: Minimum 100Mbps download, 75Mbps upload (no longer a factor in the multipliers for either consumer grade or enterprise grade compute).
- **Operating System**: Linux.
- **Cohort Membership**: Join the Render Network Foundation's enterprise cohort (in addition to RNP-019 consumer cohort), with initial onboarding limited to verified operators.
- **Uptime and Performance**: Maintain uptime and pass performance tests (e.g., benchmarked TFLOPS/TOPS, latency).

- **Multi-Node Support**: Support clustering (e.g., 8–128 GPUs) using Infiniband (NVIDIA) or Infinity Fabric (AMD) for distributed tasks like video model training.

Upload / download bandwidth is a minimum requirement and no longer factors into the Specification Multiplier (removed from RNP-019 multipliers for both consumer and enterprise compute nodes).

Job Allocation

Initially retains RNP-019 mechanics driven by customer-specified GPU types subject to availability, and evolving to an allocation model with larger customer choice - with criteria such as hardware type, price, availability and node reputation:

- **Hardware Specification**: Customers select hardware based on workload (e.g., H200 for video model training, L4 for inference, A100 for scientific simulations).
- **Price:** as determined by a dynamic auction model.
- **Performance Scoring**: Customers have the option to prioritise nodes with high uptime and performance metrics.
- **Uptime**: Necessarily being online to receive a job.
- **Multi-Node Clustering**: Support large-scale jobs via technologies like SWARM parallelism or DiLoCo, critical for video generation models.

Reward Structure

In the near term, RNP-021 expands RNP-019's rewards in a unified compute emissions pool for both consumer and enterprise compute nodes, separate from the rendering pool. We suggest a dual approach that allows for idle compute nodes to be added to the Network in the traditional way, and the ability to procure GPU time and scale enterprise-grade GPUs (e.g., NVIDIA H200, H100, or equivalents) up or down through block rentals, based on user demand. Rewards for both approaches would come from the same allocation and caps apply to the combined supply that is likely to be a mix of both traditional and block rental.

Once a dynamic auction process is built out, allow nodes to set their own asking price for rewards within the existing BME systems.

Rewards for Nodes

Pre dynamic auction model implementation, availability rewards are consistent with RNP-019 formulas.

- **Availability Rewards**: 2 RENDER/epoch per compute node, prorated by uptime (e.g., 99.5% = 1.99 RENDER), epochs aligned weekly (168 hours). No change from RNP-019; applies uniformly. RNP-021 gives the Foundation authority to reduce availability rewards as

job rewards increase, or to increase them up to 3x as per RNP-019 if market forces require it.

- **Job Rewards**: To adjust for market shifts, increase the current baseline rewards level from 10 RENDER/epoch to 25 RENDER/epoch for RTX 4090 at 100% utilization, scaled by Specification Multiplier and time worked (this adjustment falls within the remit of RNP-019 to "increase these rewards by up to 3x". Formula: 25 RENDER × Multiplier × (Hours Worked / 168).
    - **Multiplier for Enterprise GPUs**: Scaled from RTX 4090 baseline (Multiplier = 1). For H200: 5.0 (reflecting ~5x market pricing and compute value). Other examples:
        - H100: 4.0 (~80% of H200 TFLOPS).
        - A100: 3.5 (~70% of H200).
        - MI300X: 4.0 (~80% of H200).
        - L40: 2.5 (~50% of H200).
        - L4: 1.5 (~30% of H200).
        - T4: 1.2 (~20% of H200).
    - Example: H200, 100 hours/epoch = 25 × 5.0 × (100/168) ≈ 74.40 RENDER/epoch.
- **Unified Compute Emissions Pool**: RNP-019 RXT 4090 baseline of 10 RENDER increased to 25 RENDER and rewards pool expanded to accommodate enterprise cohort without diluting consumer rewards. RNP-019 baseline assumes 4,500–9,000 RENDER/month for 100 RTX 4090s.
- **Funding rewards:** funding of rewards to come from the balance of current emissions allocated to node rewards under RNP-018, plus the ability to supplement rewards by drawing from unallocated Grant emissions.
- **Node-Driven Rewards:** move to an open market model with dynamic auction pricing system that allows compute nodes to set their own asking price for rewards once the capability to run this is built. Job allocation continues to allow customers the choice of hardware, or run on what hardware is available as per their preference.

RENDER Rewards Calculation for "Traditional" Compute Nodes

Extending RNP-019 baseline (RTX 4090, Multiplier = 1):

- **Per Epoch Example**:
    - RTX 4090 (100 hours): 25 × 1 × (100/168) ≈ 14.88 RENDER (job) + 1.99 (availability) = 16.87 total.
    - H200 (100 hours): 25 × 5 × (100/168) ≈ 74.40 RENDER (job) + 1.99 (availability) = 76.39 total.
- **Monthly Example** (4.33 epochs/month):
    - RTX 4090 (100 hours/epoch): ~73.05 job + 8.62 availability = ~81.67 total.
    - H200 (100 hours/epoch): 322.15 job + 8.62 availability = ~330.77 total.
- **Block rentals (balance of up to 1,200 H200-Equivalent)**: Subject to a combined cap of 100,000 RENDER/month 100% backed by utilization and driven by market rates when applied to block rentals.
- **Note:** Traditional enterprise node operators are eligible for this work if they offer sufficient scale.

Enterprise-Grade GPU Procurement and Capacity Management

To enable the Network service customers who wish to block book more capacity than is available on traditional nodes, RNP-021 empowers the Render Network Foundation to procure block purchases of enterprise-grade GPU time (e.g., NVIDIA H200, H100, or equivalents) when approached by customers for this service. The Foundation would procure 1:1 matches with customer requests. For example: Customer requests 200 x H100s for one week. Render Network mirrors this by securing 200 x H100s for one week and makes them available to the customer.

This mechanism addresses the inherent challenges in matching blocks of demand with predictable supply of these specialized machines. Enterprise GPUs like the H200 are optimized for continuous, high-utilization environments - such as 24/7 AI training for video generation models requiring sustained throughput and massive VRAM - but their cost and scarcity can hinder decentralized network growth, as individual compute node operators face barriers to acquisition and deployment.

Via on-demand procurement of blocks of GPU time, the Foundation can secure compute supply as customers need, without paying the overhead. At scale, there may be opportunities for negotiated bulk rates (e.g., via direct OEM partnerships or authorized distributors), bypassing retail bottlenecks. Procured compute node capacity will not receive availability rewards (2-6 RENDER/epoch prorated by uptime, as in RNP-019), eliminating subsidies for idle time and aligning incentives purely with utilization. Instead, operators will be compensated at the Foundation's block purchase cost at spot rates that are significantly lower than on-demand rates, assuming a full-time availability commitment.

For marketing purposes and to maintain competitiveness against centralized providers (e.g., Runpod or Vast.ai rates), the Foundation may subsidize user costs through targeted grants from the Render Grants Pool.

These grants would be connected to account funding, for example, the customer adds $1,000 credit to their account by credit card, and the Foundation adds a matching grant as credit on the account for $200, reducing effective user rates and increasing competitiveness.

Traditional Enterprise Node Operators may compete to fill these block orders subject to any necessary operational requirements.

Objectively, this model enhances network resilience by accessing additional capacity when demand calls for it, and drives adoption - evidenced by similar strategies in other projects, where spot provisioning increased utilization by 30% without inflating token emissions. Ultimately, it positions Render as a cost-effective, decentralized alternative, capable of scaling H200-equivalent compute while preserving economic sustainability.

The effect is to allow the network to scale on an as needed basis with the aim of matching incoming demand with supply procurement. This limits the need for bloating capacity and related availability rewards when the revenue to offset is not there.

RENDER Rewards Calculation for Block Rental Compute Nodes

Block rentals would be secured on the open market, likely in fiat. This proposal creates the ability to scale up and down as needed, up to a combined cap of 100,000 RENDER/month for the remainder of 2025, for up to 1,200 H200-equivalent compute, depending on pricing. Note, the combined cap is for traditional node rewards and block rental contracts together.

Allocation for BME Year 3 (~2026) in excess of 100,000 RENDER per month to be confirmed in a future RNP.

Indicative Pricing

Pre dynamic auction model, customer pricing is based on market, competitor pricing and performance benchmarks as per RNP-019. Pricing for new models follows the methodology above, benchmarking against H200 and competitor rates. For example:

- Global Compute Index pricing: https://globalcomputeindex.com/
- Runpod Pricing: https://runpod.io/pricing
- Vast.ai Pricing: https://www.vast.ai/pricing

Post roll-out of a dynamic pricing / auction model, customer price will be driven by node asking price, grossed up to cover the Network operator fee of 5%.

Ability for the Foundation to match customer deposits with granting credit on customer accounts as a marketing / customer acquisition cost.

Example - traditional node:

Node with an H200 asks $1.90 per hour equivalent in rewards. Customer price would be $1.90 x (1+5/95) = $2.00 per hour.

Usage for 10 hours = $20.00; burn $19.00 (95%) in RENDER; $1.00 (5%) Network operator fee. Burns for compute to be handled in a separate burn wallet from rendering jobs.

Ability for the Foundation to match customer deposits with granting credit on customer accounts as a marketing / customer acquisition cost.

Note: Node rewards are distributed in RENDER each epoch.

Example - block purchased compute:

The customer needs 100 x H200s for 12 hours.

The most attractive GPU supplier asks $1.90 per hour. Customer price would be $1.90 x (1+5/95) = $2.00 per hour, or $2 x 12 x 100 $2,400 total.

Customer funds their account for $2,000. Foundation provides a matching grant (for example) $400 in compute credit, effectively discounting 200 compute hours on an H200

Customer total account balance = $2,400

Customer effective rate per hour = $2,000 / 12 / 100 = $1.67

Usage for 12 hours x 100 H200s = $2,400.00; burn $2,280.00 equivalent (95%) in RENDER; $120.00 (5%) Network operator fee. Burns for compute to be handled in a separate burn wallet from rendering jobs.

Note: GPU block rental provider rewards are distributed based on contract terms.

Rationale

- **Expansion of RNP-019**: Directly builds on consumer-grade compute by adding enterprise hardware, enabling significant capacity growth (1,200 H200-equivalent vs. RNP-019's initial cohort) for larger workloads.
- **Market Opportunity**: The projected $353 billion data center GPU market by 2030 underscores demand for high-VRAM GPUs to handle video/image models requiring 10–100x compute over LLMs. https://www.marketsandmarkets.com
- **Provider Interest**: Enterprise providers seek decentralized deployment; RNP-021 facilitates this via flexible compute node hardware.
- **Improved Capability**: Supports specialized jobs with enterprise GPUs, complementing RNP-019's consumer focus.
- **Unified Rewards Pool**: Streamlines emissions, with 5x scaling for enterprise to reflect value, without altering RNP-019 formulas.
- **Future-Proofing**: Dynamic inclusion prepares for innovations like Groq LPUs.
- **Competitive Edge**: Positions Render against AWS/CoreWeave with decentralized, scalable compute.
- **Dynamic Auction Model:** Evolve to a node-driven pricing model that allows for growth of the platform.

Impact

- **Node Operators**: Consumer compute node rewards increased 2.5x from RNP-019 for job rewards (availability rewards unchanged), enterprise compute nodes earn proportional to their compute power and market demand RNP-019, incentivizing high-end hardware.
- **Compute Clients**: Access up to ~1,200 H200-equivalent compute for demanding tasks like video generation.
- **Render Network**: Scales to enterprise level, enhancing RNP-019's compute ecosystem. Customer job revenue will follow the BME burn mechanism, with a modification on RNP-019 to allow the ability for the Network to burn RENDER directly for Compute Subnet jobs.

Potential Drawbacks

This RNP to trial enterprise compute potentially draws on a fairly large number of RENDER. The risk is mitigated by continuous monitoring and matching the speed of scaling to the speed of customer adoption so that rewards scale up and down with revenues (other than availability rewards).

Implementation Plan

1. Generate market intelligence on Enterprise GPU supply and demand through market outreach that can help inform parameters on pricing, availability, and supply.
2. Update pricing and multipliers for enterprise GPUs; benchmark new models.
3. Launch enterprise cohort, certify compute nodes for up to 1,200 H200-equivalent capacity.
4. Integrate jobs and expand pool if demand calls for it (up to ~100,000 RENDER/month for 2025).
5. Migrate to a dynamic auction process for supply and demand.
6. Evaluate and adjust based on utilization.

Community Discussion

Join #rnp-021 on Discord.

Additional References

- RNP-019: https://github.com/rndr-network/RNPs/blob/main/RNP-019.md
- Data Center GPU Market: https://www.marketsandmarkets.com/Market-Reports/data-center-gpu-market-44671175.html