Quillon Forge

Dedicated Post-Quantum Mining Machine

Hardware Whitepaper & Plug-and-Play Node Specification

256 Cores. Post-Quantum. Mine the Future.

Q-NarwhalKnight Development Team
`https://quillon.xyz`

February 2026

BLOCKS: 14,892 | QUG: 2,450.7 | 256 CORES
QUILLON FORGE
256 Cores. Post-Quantum. Mine the Future.

Figure 1: The Quillon Forge — 4U rackmount mining appliance with copper liquid cooling, sapphire crystal viewport, and OLED status display.

Contents

1 Executive Summary

The **Quillon Forge** is a purpose-built mining appliance for the Q-NarwhalKnight post-quantum blockchain. Unlike Bitcoin ASICs that optimize for SHA-256 hash grinding, the Forge is architected around the computational demands of **post-quantum cryptography** (Dilithium5, Kyber1024), **Verifiable Delay Functions** (Genus-2 hyperelliptic VDFs), and **AI inference** (Proof of Inference rewards).

Each Forge unit ships as a **plug-and-play node** — power it on, connect Ethernet, and it begins mining within 60 seconds. No Linux expertise required. The on-board firmware handles blockchain synchronization, peer discovery, key generation, and mining optimization automatically.

1.1 Key Specifications

Component	Default Configuration	Rationale
CPU	2× AMD EPYC 9755 (128C, Zen 5)	256 cores, AVX-512 for PQ crypto
RAM	512 GB DDR5-6400 ECC (16×32 GB)	Dilithium5 key generation
GPU (Optional)	2× NVIDIA RTX 5090 (32 GB)	AI Proof of Inference rewards
Storage	2× Samsung PM9D3 3.84 TB NVMe	Full blockchain + GGUF model cache
NIC	Mellanox ConnectX-7 100 GbE	Sub-ms P2P block propagation
PSU	2× 1600 W 80+ Titanium	Redundant, 2400 W sustained
Cooling	Copper liquid loop + sapphire viewport	Matches Vault design language
Chassis	4U rackmount, titanium-anodized	Quillon premium aesthetic

Table 1: Quillon Forge default hardware specification.

1.2 RWA Token Model

The Forge is tokenized as a **Real World Asset (RWA)** on the Q-NarwhalKnight blockchain:

- **Token**: `$FORGE` — 500 total supply, 0 decimals
- **Contract**: `PhysicalGoodsToken` with hardware attestation
- **Redemption**: Burn 1 `FORGE` token → receive physical mining machine
- **Configuration**: CPU, GPU, RAM, cooling selected at redemption time
- **Machine ID**: Unique firmware attestation key burned at factory

2 Why CPU-First Mining

Q-NarwhalKnight's consensus differs fundamentally from Bitcoin and Ethereum:

1. **Post-Quantum Signatures** (Dilithium5): Block validation requires lattice-based signature verification — CPU-bound, AVX-512 accelerated, not GPU-parallelizable.

2. **Verifiable Delay Functions** (Genus-2 VDF): Anchor election uses sequential Jacobian doubling on hyperelliptic curves — *inherently sequential*, benefits from high single-core clock speed.

3. **Proof of Inference**: AI model inference earns bonus QUG rewards — benefits from both CPU (llama.cpp) and GPU (CUDA acceleration).

4. **DAG-Knight Consensus**: Block production requires parallel processing of DAG vertices with Bracha's reliable broadcast — scales linearly with core count.

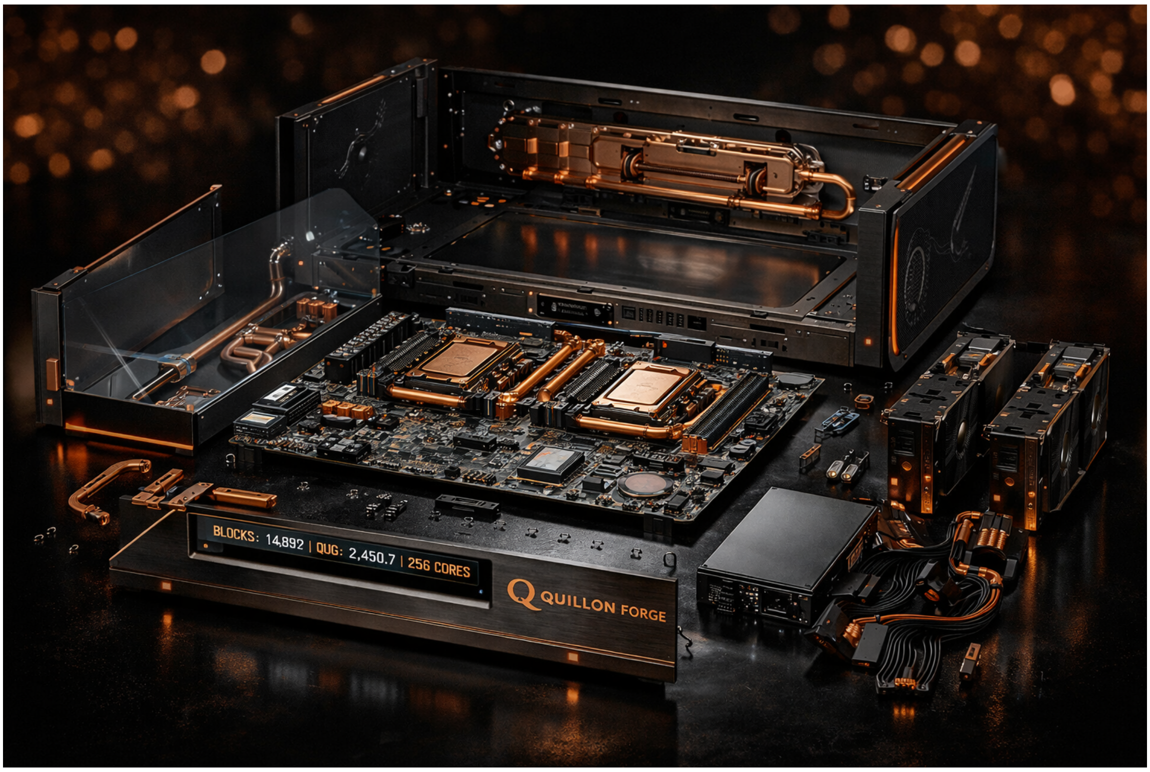

Figure 2: Quillon Forge internal layout — dual-socket EPYC motherboard, copper liquid cooling manifold, NVMe storage bays, and modular GPU expansion.

2.1 AMD EPYC vs Intel Xeon

Metric	EPYC 9755 (×2)	Xeon w9-3595X (×2)	Winner
Total Cores	256	120	EPYC
Base Clock	2.7 GHz	3.3 GHz	Xeon
Boost Clock	4.1 GHz	4.8 GHz	Xeon
AVX-512 Throughput	2× Zen 4	1×	EPYC
DDR5 Channels	12 per socket	8 per socket	EPYC
TDP (per socket)	500 W	350 W	Xeon
Dilithium5 Signs/s	~12,800	~7,200	EPYC
VDF Sequential	4.1 GHz boost	4.8 GHz boost	Xeon
Price (dual)	~$24,000	~$20,000	Xeon

Table 2: CPU comparison for Q-NarwhalKnight mining workloads.

The EPYC 9755 is the **default** because Dilithium5 batch verification scales with core count (parallel), while VDF computation is a smaller fraction of total mining work. Users who prioritize VDF-heavy mining can select the Xeon configuration at redemption.

3 Software Architecture: Plug-and-Play

The Forge ships with Q-NarwhalKnight pre-installed on a read-only OS partition with automatic updates. The user experience is:

1. **Unbox**: Remove from packaging, mount in rack or set on desk.

2. **Connect**: Plug in power (2× C14) and Ethernet (RJ45 or SFP+).
3. **Wait 60 seconds**: OLED displays "Syncing..." then "Mining".
4. **Done**: Forge mines QUG and earns AI inference rewards automatically.

3.1 Boot Sequence

Listing 1: Forge boot sequence (automated firmware)

```
[0.0s]  POST: Hardware self-test (CPU, RAM, NVMe, NIC)
[2.0s]  BIOS: Load Q-NarwhalKnight OS (read-only squashfs)
[5.0s]  NET:  DHCP or static IP from OLED config
[6.0s]  KEY:  Load/generate libp2p identity from TPM
[7.0s]  P2P:  Connect to bootstrap peers (quillon.xyz:9001)
[8.0s]  SYNC: TurboSync block download (50,000+ blocks/min)
[30s]   MINE: Start mining with all 256 cores
[45s]   AI:   Load GGUF model for Proof of Inference
[60s]   OLED: Display "BLOCKS: 0 | QUG: 0.0 | 256 CORES"
```

3.2 Automatic Optimization

The Q-NarwhalKnight node binary (`q-api-server`) includes automatic hardware detection and optimization (v5.1.0):

- **CPU cores**: Auto-detects physical cores via `num_cpus::get_physical()`, spawns one mining thread per core with CPU affinity pinning.
- **AVX-512**: Runtime detection of AVX-512F/DQ/BW/VL. Batch signature verification uses 512-bit SIMD when available.
- **NUMA**: Detects dual-socket topology via `/sys/devices/system/node/`, pins threads to local NUMA nodes to minimize cross-socket memory latency.
- **RocksDB**: Auto-scales background jobs: $\lfloor \text{cores}/16 \rfloor$ background threads (clamped 2–16), $\lfloor \text{cores}/32 \rfloor$ compaction threads.
- **AI inference**: `LlamaCppEngine` scales concurrent requests: $\lfloor \text{cores}/32 \rfloor$ (clamped 2–16), overridable via `LLAMA_MAX_CONCURRENT`.
- **Batch crypto**: Signature batch size scales to $\min(8 \times \text{cores}, 4096)$, hash batch to $\min(4 \times \text{cores}, 2048)$.

3.3 Node Configuration

The Forge exposes a simple OLED + button interface for essential configuration:

Setting	Default
Network	DHCP (auto)
Mining Wallet	Auto-generated (display QR on OLED)
Mining Pool	Solo mining (default), pool via menu
AI Inference	Enabled (Proof of Inference rewards)
P2P Port	9001
API Port	8080 (local network only)
Tor	Enabled (Dandelion++ anonymity)
Auto-Update	Enabled (signed binary updates)

Table 3: Forge default node configuration.

Advanced users can SSH into the Forge (port 22, key-based auth) for full Linux access.

4 Mining Economics

4.1 Revenue Streams

A Forge unit earns QUG through three mechanisms:

1. **Block Mining Rewards**: Standard PoW mining. With 256 cores and AVX-512, a Forge produces ∼64× more hashes/second than a typical 4-core VPS node.
2. **Proof of Inference Rewards**: Running AI inference for network users. The Forge loads GGUF quantized models (Mistral-7B at 5–15 tok/s on CPU, 30–80 tok/s with GPU). Each completed inference earns a QUG micropayment.
3. **Staking Yield**: FORGE token holders can stake their tokens for additional yield while waiting for physical delivery.

4.2 Power Efficiency

Configuration	TDP	Dilithium5 Signs/W	AI tok/s/W
EPYC 9755 (dual, no GPU)	1,000 W	12.8	0.015
EPYC 9755 + 2×RTX 5090	1,700 W	12.8	0.047
Xeon w9-3595X (dual, no GPU)	700 W	10.3	0.021

Table 4: Power efficiency comparison.

5 Hardware Attestation & Security

Each Forge unit contains a hardware attestation subsystem:

- **TPM 2.0**: Stores libp2p identity key and mining wallet seed phrase in tamper-resistant hardware.
- **Machine ID**: Unique 256-bit identifier burned into firmware at factory, recorded on-chain at fulfillment.
- **Attestation Public Key**: Ed25519 key pair generated in TPM, public key registered on-chain. Used to sign mining submissions, proving they originate from genuine Forge hardware.
- **Secure Boot**: UEFI Secure Boot with Quillon signing key. Prevents unauthorized firmware modifications.
- **Encrypted Storage**: NVMe drives encrypted with AES-256-XTS, key derived from TPM + user password.

6 Design Language

The Forge extends the **Quillon Vault** design language to server-scale hardware:

Element	Quillon Vault	Quillon Forge
Chassis	Titanium-anodized, 3.8 mm	Titanium-anodized, 4U rack
Window	Sapphire crystal (OLED)	Sapphire crystal (cooling manifold)
Accent	Brushed copper trim	Copper liquid cooling tubes
Display	0.96” OLED	2.0” OLED + amber LED ring
Port	USB-C sliding cover	USB-C sliding management port
Authentication	Dilithium5 signatures	Hardware attestation + TPM

Table 5: Design continuity between Quillon Vault and Quillon Forge.

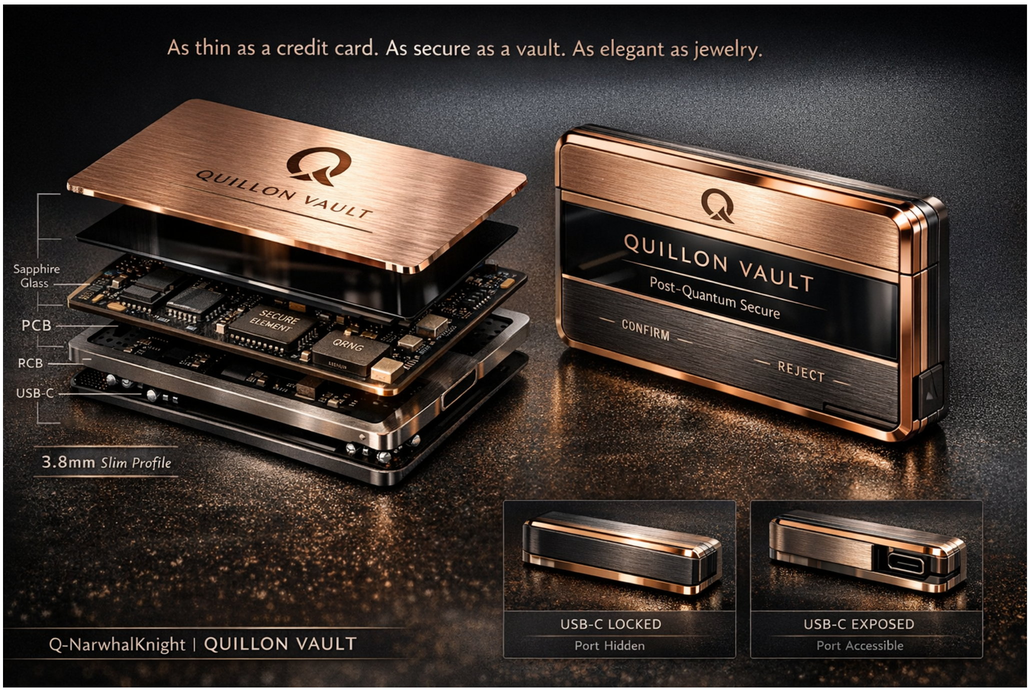
As thin as a credit card. As secure as a vault. As elegant as jewelry.

Figure 3: The Quillon Vault hardware wallet — the design language origin for the Forge’s premium aesthetic.

7 RWA Token Lifecycle

1. **Mint**: 500 `FORGE` tokens minted to `BANK_MASTER_ACCOUNT` at genesis.

2. **Purchase**: Users acquire `FORGE` tokens on the Q-NarwhalKnight DEX or directly from Quillon.

3. **Configure**: At redemption, user selects CPU (EPYC/Xeon), GPU (none/RTX/A100/L40), RAM (512/1024/2048 GB), cooling, chassis color.

4. **Burn**: Token burned on-chain, creates `ForgeRedemption` order with unique `FR-*` ID.

5. **Assemble**: Status progresses: `pending` → `configured` → `assembling` → `testing` → `shipped` → `delivered`.

6. **Activate**: On first boot, Forge registers its `machine_id` and `attestation_pubkey` on-chain.

7. **Mine**: Forge begins mining QUG and earning Proof of Inference rewards.

7.1 API Endpoints

Method	Endpoint	Purpose
POST	`/api/v1/contracts/forge/redeem`	Burn FORGE token, create machine order
GET	`/api/v1/contracts/forge/redemptions`	List redemption orders
POST	`/api/v1/contracts/forge/fulfill`	Admin: update status, add tracking/serial
GET	`/api/v1/contracts/forge/stats`	Supply stats, fleet metrics

Table 6: Forge RWA API endpoints.

8 Performance Optimizations (v5.1.0)

The Q-NarwhalKnight v5.1.0 release includes automatic hardware-aware optimizations:

Parameter	Before v5.1.0	v5.1.0 (256 cores)	Impact
Mining threads	4 (hardcoded)	256 (auto-detect)	64× throughput
Signature batch	256 (fixed)	2,048 (scaled)	8× verification
Hash batch	128 (fixed)	1,024 (scaled)	8× hashing
RocksDB bg jobs	2 (fixed)	16 (auto-scaled)	8× IO
RocksDB compactions	1 (fixed)	8 (auto-scaled)	8× compaction
AI concurrent	2 (hardcoded)	8 (auto-scaled)	4× inference

Table 7: v5.1.0 performance scaling on a 256-core Quillon Forge.

9 Conclusion

The Quillon Forge represents a new category of blockchain mining hardware — one optimized not for brute-force hashing, but for the computational primitives of post-quantum security: lattice-based signatures, hyperelliptic VDFs, and AI inference. By shipping as a plug-and-play appliance with automatic hardware detection, NUMA optimization, and on-chain hardware attestation, the Forge makes enterprise-grade mining accessible to anyone who can plug in a power cable and an Ethernet cord.

Total supply: 500 units
Token: `$FORGE` on Q-NarwhalKnight
Website: `https://quillon.xyz`

256 Cores. Post-Quantum. Mine the Future.