QUILLON FORGE — Full Manufacturing Pipeline

February 2026

Contents

QUILLON FORGE — Full Manufacturing Pipeline

From Whitepaper to Plug-and-Play Mining Machine

Status: Planning Phase **Estimated Timeline:** 12–18 months **Estimated Total Investment:** $800K–2.5M (for first 500 units) **Target Retail Price:** $18,500 (Base) / $24,500 (GPU-equipped) / $32,000 (Full AI Rig) **Target BOM Cost:** ~$8,200/unit @ 500 volume **RWA Token:** $FORGE — 500 total supply, 0 decimals, burn-to-redeem

---

PHASE 0: DESIGN & DOCUMENTATION (COMPLETE)

Deliverables Completed

☒ Product concept and hardware specification
☒ Whitepaper with full hardware architecture (`papers/quillon-forge-whitepaper.pdf`)
☒ Product renders / concept images (`docs/announcements/forge1.png`, `forge2.png`)
☒ Component BOM with specific part numbers
☒ Q-NarwhalKnight blockchain integration ($FORGE RWA token)
☒ Backend API endpoints (`/api/v1/contracts/forge/*`)
☒ Performance auto-scaling software (v5.1.0)
☒ Quillon Vault design language reference (`docs/vault.png`)

Deliverables — Claude Can Produce Next

☐ Detailed mechanical drawings (4U chassis, cooling manifold, airflow simulation)
☐ Electrical schematic for OLED status panel + LED ring controller
☐ Custom PCB for management controller (BMC — Baseboard Management Controller)
☐ Firmware for BMC (auto-boot, health monitoring, OLED driver, remote management)
☐ TPM attestation firmware (machine-ID enrollment, remote attestation protocol)
☐ Factory test suite (stress test, cooling validation, burn-in scripts)
☐ Packaging specification (crate, foam insert, accessory kit)
☐ Assembly line work instructions (illustrated step-by-step)
☐ QC pass/fail criteria document

PHASE 1: SYSTEM ARCHITECTURE & COMPONENT SELECTION

Timeline: Weeks 1–6 **Budget:** $15–30K

What Happens

1. **Motherboard Selection** — Dual-socket SP5 (EPYC 9005-series) platforms:

Motherboard	Form Factor	Slots	NVMe	NICs	Price
Supermicro H13DSH	EATX	24 DIMM	4× M.2	2× 25GbE	~$1,200
ASRock Rack ROMED8HM3	EATX	16 DIMM	2× M.2	2× 10GbE	~$800
Tyan S8253	SSI EEB	24 DIMM	4× M.2	2× 10GbE	~$1,100

   Selected: Supermicro H13DSH — Most expansion, best IPMI/BMC support, proven in HPC.

2. **CPU Validation** — Confirm EPYC 9755 availability and AVX-512 performance:
   - AMD EPYC 9755 (Turin, Zen 5, 128C/256T, 2.7/4.1 GHz, 500W TDP)
   - Verify AVX-512 `vpclmulqdq` instruction support for Dilithium5 NTT
   - Benchmark: `cargo bench --package q-crypto-simd` on dual-socket EPYC eval kit
   - Alternative: EPYC 9654 (96C, Zen 4, $5,500 — budget option)

3. **Memory Configuration** — DDR5 ECC for crypto workloads:
   - 16× Samsung M321R4GA0BB0-CQK (32GB DDR5-4800 ECC RDIMM)
   - Total: 512GB across 12 channels (6 per socket)
   - Dilithium5 key gen needs ~128MB working set per thread
   - 512GB allows 256 concurrent mining threads without swap pressure

4. **Storage Selection** — NVMe for blockchain + AI models:
   - 2× Samsung PM9D3 3.84TB U.2 NVMe (enterprise, 1 DWPD)
   - RAID-1 mirror for blockchain integrity (no data loss on single drive failure)
   - GGUF model cache: ~50GB for Mistral-7B + Qwen-VL-8B quantized models
   - RocksDB: auto-scaled with v5.1.0 (16 bg jobs, 8 compaction threads on 256 cores)

5. **GPU Evaluation** (optional configurations):

GPU	VRAM	TDP	AI tok/s	Price	Config Name
None	—	0W	CPU only	$0	`none`
2× RTX 5090	2×32GB	2×575W	~80 tok/s	~$4,000	`rtx-5090-dual`
4× RTX 5090	4×32GB	4×575W	~160 tok/s	~$8,000	`rtx-5090-quad`
2× A100 80GB	2×80GB	2×300W	~120 tok/s	~$20,000	`a100-dual`
4× L40 48GB	4×48GB	4×300W	~200 tok/s	~$28,000	`l40-quad`

6. **Network Interface** — Low-latency P2P:
   - Mellanox ConnectX-7 100GbE (2-port QSFP56) — $350
   - Fallback: Intel X710-DA2 10GbE — $120
   - Sub-millisecond block propagation over dedicated P2P link

7. **Power Supply** — Redundant for 24/7 uptime:
   - 2× SuperMicro PWS-1K62A-1R (1600W 80+ Titanium, hot-swap)
   - Total capacity: 3200W, max sustained draw: ~2400W (GPU config)
   - CPU-only config: ~1200W sustained, single PSU sufficient

Who to Source From

- **Supermicro** — Motherboard, chassis tray, PSUs, IPMI management
- **AMD** — Direct EPYC 9755 allocation (HPC partner program)
- **Samsung Semiconductor** — NVMe drives (enterprise channel)
- **Mellanox/NVIDIA** — ConnectX-7 NICs, GPU allocations
- **Mouser/Digikey** — DDR5 DIMMs, cables, thermal paste, misc components

Claude's Role in Phase 1

- Write benchmark scripts for AVX-512 Dilithium5 performance validation
- Generate exact BOM spreadsheet with MPN, quantity, unit cost, supplier
- Write BIOS configuration script (SR-IOV, NUMA interleave policy, AVX-512 power limits)
- Create thermal model spreadsheet (component TDP → required airflow CFM → cooling design)

PHASE 2: MECHANICAL ENGINEERING — CHASSIS & COOLING

Timeline: Weeks 4–12 (parallel with Phase 1) **Budget:** $40–80K

What Happens

1. **Chassis Design** (SolidWorks / Fusion360):
   - 4U rackmount enclosure (19" standard, 660mm depth)
   - Titanium-anodized front panel with copper accent trim
   - Sapphire crystal viewport over cooling manifold (matches Vault design language)
   - OLED + LED ring bezel (2.0" OLED behind polycarbonate lens)
   - Front I/O: USB-C management port (sliding cover — Vault-style), OLED, 2× buttons
   - Rear I/O: 2× QSFP56/RJ45, 2× C14 power, IPMI RJ45, 4× SFP+ expansion
   - Internal: Supermicro EATX tray + GPU riser bracket zone
   - Weight target: <30kg (shippable via standard freight)
2. **Liquid Cooling System Design**:

```
FRONT (COLD SIDE)                       REAR (HOT SIDE)


  2×120mm       COPPER          2×280mm RADIATOR
  INTAKE      MANIFOLD       + 4×140mm FANS
  FANS          (CPU×2)         (exhaust to rear)




                  PUMP     EK-XTOP Revo D5 (PWM)
                  (×2      Redundant - auto-failover
                 redundt)


Loop: Pump → CPU0 cold plate → CPU1 cold plate →
      VRM heatsink → radiator → reservoir → pump

Coolant: EK-CryoFuel Clear (propylene glycol + biocide)
Tubing: 12mm copper hardline (matches aesthetic)
Fittings: EK-Quantum Torque compression (nickel-plated)
```

3. **Thermal Analysis**:
    - CPU-only: 2×500W = 1000W + VRM 50W + NVMe 30W = **1080W thermal load**
    - GPU config: +2×575W = **2230W thermal load**
    - Radiator capacity: 2×280mm = 560mm radiator = ~1200W dissipation at $\Delta T=15$°C
    - GPU config requires additional 2×360mm rear radiator (+1400W capacity)
    - Target ambient: 25°C room, CPU junction <85°C under sustained AVX-512 load
    - Airflow: 200+ CFM through radiators (4×140mm Noctua NF-A14 iPPC-3000)
4. **CNC Prototyping**:
    - Front panel: 6061 aluminum prototype first ($300), then Grade 5 Ti ($1,500)
    - Copper cooling manifold: CNC from C110 copper bar ($400–800/unit)
    - Sapphire viewport: 80mm × 40mm × 2mm, AR-coated (watch crystal supplier)
    - 3 complete chassis prototypes for fit testing
5. **LED Ring & OLED Bezel**:
    - 24× WS2812B addressable LEDs in copper ring (amber = mining, blue = syncing, red = error)
    - 2.0” SSD1309 OLED (128×64) for status display
    - Custom PCB: STM32G0 microcontroller driving OLED + LEDs + 2 buttons
    - I2C connection to BMC for status updates

Who to Hire

- **ME / Thermal Engineer** with server chassis experience: $15–30K
- **CNC fabrication**: Fictiv, Xometry, Protolabs (aluminum prototypes)
- **Titanium machining**: Precision Castparts, Weber Manufacturing (production)
- **Copper manifold**: Custom CNC shops (C110 oxygen-free copper)
- **Sapphire viewport**: Stettler Sapphire (Swiss), GT Advanced Technologies
- **Liquid cooling components**: EK Water Blocks (OEM program), Alphacool (industrial)

Claude’s Role in Phase 2

- Write thermal simulation parameters (heat source locations, CFM requirements)
- Design OLED + LED ring controller PCB schematic (STM32G0 + WS2812B + SSD1309)
- Write LED animation patterns (boot: sweep, mining: pulse, error: flash, sync: breathe)
- Spec all cooling components with exact part numbers and quantities
- Write assembly instruction document with clearance verification steps

---

PHASE 3: BASEBOARD MANAGEMENT CONTROLLER (BMC) FIRMWARE

Timeline: Weeks 6–16 **Budget:** $0 (Claude writes it) to $10–20K (if hiring for IPMI/BMC expert review)

Architecture

```
                    FORGE MANAGEMENT CONTROLLER


      STM32G0B1 MCU            Supermicro IPMI/BMC
      (OLED + LED ring)        (built into motherboard)
```

```
SSD1309 OLED            Temperature sensors
24× WS2812B LEDs           Fan speed control
2× tactile buttons         Power state management
USB-C management           Remote KVM / Serial-over-LAN
I2C slave to BMC           Redfish API (HTTPS)
```

```
           Q-NarwhalKnight Node Binary
           (q-api-server v5.1.0+)

Auto-detect: 256 cores, AVX-512, NUMA, NVMe
Mining: 256 threads (1 per physical core)
RocksDB: 16 bg jobs, 8 compaction, 4 flush
AI: LlamaCppEngine @ 8 concurrent (LLAMA_MAX_CONCURRENT)
P2P: libp2p + Tor (Dandelion++)
```

Firmware Modules (Claude Writes All)

1. **Boot Sequence Controller**
    - POST self-test: CPU, RAM, NVMe, NIC, cooling, PSU
    - OLED splash: "QUILLON FORGE" logo (copper on black)
    - DHCP / static IP acquisition (OLED displays IP)
    - Start `q-api-server` binary with auto-detected hardware flags
    - OLED display: `BLOCKS: 0 | QUG: 0.0 | 256 CORES`
    - LED ring: amber mining pulse
2. **OLED Status Display** (real-time, updated every 2 seconds)

    ```
    BLOCKS: 14,892
    QUG: 2,450.7   256 CORES
    TEMP: 72°C   FAN: 2100 RPM
    NET: 12 peers    15 MB/s
    ```

    - Page 1: Blocks, QUG balance, core count
    - Page 2: CPU temp, fan RPM, coolant temp
    - Page 3: Peer count, bandwidth, sync %
    - Page 4: AI inference requests served, tok/s
    - Button 1: Cycle pages
    - Button 2 (hold 3s): Show wallet QR code
    - Both buttons (hold 10s): Factory reset
3. **Health Monitoring Daemon**
    - CPU temperature: warn >85°C, throttle >90°C, shutdown >95°C
    - Coolant temperature: warn >45°C, shutdown >55°C
    - Pump RPM: alert if either pump fails (<100 RPM)
    - Fan RPM: alert if any fan fails, ramp remaining fans to 100%
    - NVMe health: SMART monitoring, warn on wear level >80%
    - PSU: monitor both rails, alert on single PSU loss
    - RAM: ECC error logging, alert on uncorrectable errors

4. **Auto-Update System**
    - Check `https://quillon.xyz/downloads/` for new `q-api-server` binary
    - Verify Ed25519 signature on binary (signed by Quillon release key)
    - Display update prompt on OLED: “Update v5.1.0 → v5.2.0? [YES] [NO]”
    - Apply update: stop mining → backup current binary → replace → restart
    - Rollback if new binary fails health check within 60 seconds
5. **TPM Attestation Module**
    - On first boot: generate libp2p Ed25519 identity in TPM 2.0
    - Generate attestation key pair (Ed25519) in TPM, export public key
    - Register `machine_id` + `attestation_pubkey` on-chain via Forge fulfill API
    - Sign mining submissions with TPM-held key (proves genuine Forge hardware)
    - Remote attestation: challenge-response protocol for fleet management
6. **Remote Management Interface**
    - SSH (port 22): key-based auth only, root disabled by default
    - Web dashboard (port 8443): BMC health metrics, node status, basic config
    - IPMI: Supermicro native IPMI for power cycling, serial console
    - API: `/api/v1/forge/health` endpoint with full system telemetry

Development Tools

- **Toolchain**: ARM GCC for STM32G0, Linux GCC for BMC daemon
- **Build**: CMake (STM32), Cargo (node binary)
- **Debug**: SWD/JTAG for STM32, SSH for Linux
- **Testing**: Hardware-in-the-loop test bench (thermal chambers, load generators)

PHASE 4: OPERATING SYSTEM & NODE SOFTWARE

Timeline: Weeks 8–16 **Budget:** $0 (Claude writes it) to $5K (if hiring sysadmin for hardening review)

Forge OS (Custom Linux)

```
FORGE OS (based on Debian 12 minimal)

Partition Layout:
/dev/nvme0n1p1  512MB  EFI System Partition
/dev/nvme0n1p2  4GB    rootfs (read-only squashfs)
/dev/nvme0n1p3  16GB   /var (logs, config)
/dev/md0        REST   /data (RAID-1 blockchain)

Services (systemd):
   q-api-server.service    (mining node)
   forge-bmc.service       (OLED/LED/health)
   forge-watchdog.service  (auto-restart on crash)
   forge-updater.timer     (check updates hourly)
   sshd.service            (remote access)

Kernel: 6.6 LTS with:
- CONFIG_X86_SGX=y (Intel attestation)
- CONFIG_CRYPTO_DILITHIUM=m (PQ kernel module)
- CONFIG_NUMA=y (dual-socket NUMA awareness)
```

```
  - CONFIG_CPU_FREQ_GOV_PERFORMANCE=y (no throttle)
  - CONFIG_VFIO=y (GPU passthrough for AI)
```

Plug-and-Play Boot Flow (Total: 60 seconds to mining)

```
[0.0s]   UEFI POST → Secure Boot (Quillon signing key)
[2.0s]   BIOS: Load Forge OS kernel from NVMe
[4.0s]   Kernel: Detect 256 cores, 512GB RAM, NUMA topology
[5.0s]   systemd: Start forge-bmc.service → OLED shows "BOOTING..."
[6.0s]   systemd: Start networking → DHCP or static from /var/config
[7.0s]   OLED: Display assigned IP address
[8.0s]   systemd: Start q-api-server.service
[8.5s]   Node: Load/generate libp2p identity from TPM
[9.0s]   Node: Connect to bootstrap peers (quillon.xyz:9001)
[10.0s]  Node: Begin TurboSync block download
[30.0s]  Node: Sync complete (50,000+ blocks/min with 100GbE)
[31.0s]  Node: Start mining with 256 cores (auto-detected)
[45.0s]  Node: Load GGUF AI model for Proof of Inference
[50.0s]  OLED: "BLOCKS: 0 | QUG: 0.0 | 256 CORES"
[55.0s]  LED Ring: Amber mining pulse begins
[60.0s]  Fully operational - mining + AI inference + P2P
```

OLED Configuration Menu (button-navigated)

```
MAIN MENU
  Status (default view - cycles automatically)
  Network
      DHCP (auto) ← default
      Static IP: ___.___.___.__
      DNS: ___.___.___.__
  Mining
      Wallet: [Show QR] / [Import via USB]
      Pool: Solo (default) / Pool URL: ________
      CPU Threads: AUTO (256) / Custom: ___
  AI Inference
      Enabled (default) / Disabled
      Model: Mistral-7B / Qwen-VL-8B / Custom
      GPU: Auto / CPU Only / Disabled
  Security
      SSH: Enabled / Disabled
      Tor: Enabled (default) / Disabled
      Auto-Update: Enabled (default) / Disabled
  System
      Firmware Version: v5.1.0
      Machine ID: QF-2026-XXXXX
      Reboot
      Factory Reset (hold 10s)
  About
       Serial Number
       Uptime
       quillon.xyz
```

PHASE 5: FIRST PROTOTYPE

Timeline: Weeks 16–20 **Budget:** $25–60K (3 complete units)

Bill of Materials — Single Unit

Component	Part Number	Qty	Unit Cost	Total
CPUs	AMD EPYC 9755 (128C, Zen 5)	2	$5,500	$11,000
Motherboard	Supermicro H13DSH	1	$1,200	$1,200
RAM	Samsung M321R4GA0BB0 32GB DDR5 ECC	16	$85	$1,360
NVMe	Samsung PM9D3 3.84TB U.2	2	$450	$900
NIC	Mellanox ConnectX-7 100GbE	1	$350	$350
PSU	SuperMicro PWS-1K62A-1R 1600W	2	$280	$560
CPU Cold Plates	EK-Quantum Velocity² sTR5	2	$120	$240
Radiator	EK-CoolStream PE 280	2	$65	$130
Pump	EK-XTOP Revo D5 PWM	2	$85	$170
Fans (140mm)	Noctua NF-A14 iPPC-3000	4	$28	$112
Fans (120mm intake)	Noctua NF-A12x25	2	$32	$64
Copper tubing	12mm OD C110 hardline (1m)	3	$15	$45
Fittings	EK-Quantum Torque 12mm (nickel)	16	$8	$128
Reservoir	EK-Quantum Kinetic 120 DDC	1	$95	$95
Coolant	EK-CryoFuel Clear (1L)	2	$15	$30
OLED Panel	SSD1309 2.0” 128×64 I2C	1	$8	$8
LED Ring	WS2812B strip (24 LEDs, cut)	1	$4	$4
Management PCB	Custom STM32G0B1 board	1	$25	$25
TPM Module	Infineon SLB 9670 TPM 2.0	1	$18	$18
Sapphire Viewport	80×40×2mm AR-coated	1	$35	$35
Chassis	4U titanium-anodized (CNC)	1	$800	$800
Copper Manifold	CNC C110 cooling cover	1	$300	$300
Rails	Supermicro MCP-290-00058-0N	1	$45	$45
Cables	PCIe, SATA power, USB internal	1 set	$50	$50
Thermal Paste	Thermal Grizzly Kryonaut (5g)	2	$12	$24
NVMe Tray	U.2 to NVMe adapter + bay	2	$20	$40
Packaging	Pelican-style foam case	1	$80	$80
Power Cords	C14-C13 16AWG (2m) × 2	1 set	$15	$15
Ethernet Cable	CAT6A 3m (included)	1	$8	$8
USB Drive	16GB USB-C (Forge OS installer)	1	$8	$8
			BOM Total	**$17,909**

> **Note:** Prototype pricing — volume pricing (500 units) estimated at ~**$8,200/unit** with AMD/Samsung direct allocation.

Assembly Steps (3 units, hand-assembled)

1. Mount Supermicro motherboard onto 4U chassis tray, secure with standoffs
2. Install 2× EPYC 9755 CPUs with thermal paste, mount EK cold plates
3. Install 16× DDR5 DIMMs (populate all channels — 8 per socket for 6-channel)
4. Install 2× Samsung PM9D3 NVMe drives in U.2 bays
5. Install Mellanox ConnectX-7 NIC in PCIe slot
6. Install TPM 2.0 module on motherboard TPM header
7. Route copper cooling loop: pump → CPU0 → CPU1 → VRM → radiator → reservoir
8. Install 2× 280mm radiators with 4× 140mm fans (rear exhaust)
9. Install 2× 120mm intake fans (front)
10. Fill cooling loop, bleed air (24-hour leak test)

11. Install 2× PSUs in hot-swap bays
12. Install management PCB (STM32G0) with OLED + LED ring into front bezel
13. Connect management PCB I2C cable to motherboard BMC header
14. Install sapphire viewport over cooling manifold
15. Close titanium front panel, secure with captive screws
16. Flash Forge OS to NVMe via USB installer
17. First boot — OLED shows “QUILLON FORGE”
18. Run factory self-test suite (all sensors, cooling, stress test)
19. Generate libp2p identity in TPM
20. Connect to quillon.xyz bootstrap, verify sync + mining begins

Validation Checklist

- ☐ All 256 cores detected, mining threads spawn correctly
- ☐ AVX-512 detected and used for Dilithium5 batch verification
- ☐ NUMA topology correct (128 cores per socket, local memory)
- ☐ DDR5 ECC: all 512GB detected, memtest86 passes (48-hour soak)
- ☐ Both NVMe drives healthy, RAID-1 mirror active
- ☐ 100GbE NIC link up, iperf3 90 Gbps
- ☐ Cooling: CPU <85°C under sustained AVX-512 Prime95 (2-hour test)
- ☐ Cooling: no leaks after 24-hour pressure test
- ☐ Pump redundancy: disconnect pump 1, pump 2 takes over, temps stable
- ☐ PSU redundancy: disconnect PSU 1, system stays running
- ☐ OLED displays all 4 status pages correctly
- ☐ LED ring: all 24 LEDs addressable, color-correct
- ☐ TPM: identity generated, attestation key exported
- ☐ P2P: connects to bootstrap, syncs blocks, starts mining within 60s
- ☐ AI inference: loads GGUF model, responds to test prompt
- ☐ Power draw: <1300W (CPU-only), <2500W (GPU config) at wall
- ☐ Noise: <55 dBA at 1m (mining load, fans at auto)
- ☐ Weight: <30kg assembled

PHASE 6: BURN-IN TESTING & SECURITY AUDIT

Timeline: Weeks 20–28 **Budget:** $30–80K

Burn-In Testing (Each prototype, 168 hours = 7 days continuous)

1. **CPU Stress**: `stress-ng --cpu 256 --cpu-method matrixprod` for 168 hours
2. **AVX-512 Stress**: Custom Dilithium5 batch verification loop at 100% for 168 hours
3. **Memory**: memtest86+ extended test (48 hours), then application-level (120 hours)
4. **Storage**: FIO random 4K write/read at queue depth 32 for 168 hours
5. **Cooling**: Ambient temperature cycling (15°C → 35°C → 15°C) every 6 hours
6. **Network**: iperf3 bidirectional 100Gbps for 24 hours, P2P block relay for 144 hours
7. **Mining Soak**: Actual Q-NarwhalKnight mining for 168 hours, verify rewards accumulate
8. **Power Cycle**: 50× graceful shutdown/boot cycles, verify clean startup every time
9. **PSU Failover**: Yank PSU 1 under load 10 times, verify zero downtime
10. **Pump Failover**: Disable pump 1 under load, verify temp stabilizes on pump 2

Security Audit ($15–40K)

Firms to hire: - **NCC Group** (UK/US) — Embedded systems + crypto implementation audit - **Trail of Bits** (NYC) — Firmware + node binary security review - **Cure53** (Berlin) — Network security, remote management attack surface

What they test: - BMC attack surface: can IPMI be exploited to gain root? - TPM key extraction: side-channel analysis on attestation key - Firmware update chain: can unsigned firmware be loaded? - SSH hardening: default credentials, key management - Network exposure: scan all ports, verify only intended services exposed - Node binary: memory safety (Rust — mostly covered), API injection - Physical: can the chassis be opened without breaking tamper seal?

Regulatory Certification ($8–15K)

- **FCC Part 15 Class A** (US) — Intentional/unintentional radiator, required for commercial sale
- **CE Mark** (EU) — EMC Directive 2014/30/EU, Low Voltage Directive 2014/35/EU
- **UL/CSA** — Safety certification for server power supplies
- **RoHS** — Restriction of Hazardous Substances (EU)
- **Test labs**: TUV, Intertek, SGS, UL
- Timeline: 6–8 weeks after submission

PHASE 7: PRODUCTION

Timeline: Weeks 28–52 **Budget:** $500K–1.8M for first 500 units

Tooling ($30–60K)

- CNC fixtures for titanium front panel: $8–15K
- CNC fixtures for copper cooling manifold: $5–10K
- Custom assembly jigs (motherboard + cooling loop): $5–10K
- Factory test bench (automated): $8–15K
- Packaging foam mold (Pelican-style): $3–5K

Component Sourcing (Lead Times)

Component	Supplier	Lead Time	MOQ
AMD EPYC 9755	AMD Direct (HPC program)	16–20 weeks	100
Supermicro H13DSH	Supermicro Direct	8–12 weeks	50
Samsung DDR5 32GB ECC	Samsung Semiconductor	8 weeks	1,000
Samsung PM9D3 3.84TB	Samsung Direct	10 weeks	200
Mellanox ConnectX-7	NVIDIA/Mellanox Direct	12 weeks	100
SuperMicro PSU 1600W	Supermicro Direct	8 weeks	100
EK Water Blocks (all)	EK OEM Program	6–8 weeks	100
Noctua fans	Noctua OEM	4–6 weeks	500
Sapphire viewport	GT Advanced Technologies	10 weeks	200
Ti-6Al-4V front panels	CNC batch (Precision Castparts)	12 weeks	100
Copper manifolds	CNC batch (custom shop)	8 weeks	100
STM32G0B1 mgmt PCB	JLCPCB + assembly	3 weeks	500
TPM 2.0 modules	Infineon/Mouser	6 weeks	500
Chassis (4U body)	Sheet metal + powder coat	8 weeks	100

Critical path: AMD EPYC 9755 at 16–20 weeks. Order first, secure allocation.

Contract Assembly Partners

- **Supermicro Integration Services** (San Jose, CA) — They build custom servers to spec, highest quality, familiar with their own motherboards
- **Penguin Computing** (Fremont, CA) — HPC system integrator, experienced with liquid cooling
- **Exxact Corporation** (Fremont, CA) — GPU server specialist, custom liquid cooling experience
- **Puget Systems** (Auburn, WA) — Boutique system integrator, excellent QC
- **In-house** — Small team (3–5 people) at Quillon facility if volume justifies

Per-Unit Production Cost @ 500 units (Target)

Item	Cost
BOM (all components, volume)	$8,200
Assembly labor (8 hours @ $50/hr)	$400
Liquid cooling integration	$300
Burn-in testing (48 hours, electricity)	$50
Forge OS installation + provisioning	$50
Quality control + cosmetic inspection	$50
Packaging (Pelican case, accessories)	$120
Total per unit	**~$9,170**

Per-Unit Production Cost @ 2,000 units (Scale)

Item	Cost
BOM (volume pricing, AMD allocation)	$6,800
Assembly (semi-automated line)	$250
Liquid cooling	$200
Burn-in + QC	$60
OS + provisioning	$30
Packaging	$100
Total per unit	**~$7,440**

Quality Control (Every Unit)

1. **Automated POST**: BIOS self-test, all 256 cores, 512GB RAM, NVMe, NIC
2. **Thermal validation**: 30-minute stress test, verify CPU <85°C, no leaks
3. **Cooling pressure test**: 24-hour static pressure hold (no drips)
4. **Network test**: iperf3 link speed 90 Gbps
5. **Mining test**: 1-hour actual mining, verify block submissions
6. **AI inference test**: Load model, serve 10 test prompts, verify output
7. **OLED + LED test**: All pixels, all LED colors, button response
8. **Power measurement**: Verify <1300W (no GPU) or <2500W (GPU) at wall
9. **Noise measurement**: <55 dBA at 1m under load
10. **Cosmetic inspection**: Titanium finish, sapphire window, copper tubing
11. **TPM enrollment**: Generate identity, export attestation key
12. **Pack and seal**: Serial number label, tamper-evident seal, packing list

---

PHASE 8: LAUNCH

Timeline from start: ~12–18 months **First batch:** 500 units (matching $FORGE token supply)

Pricing

Configuration	Retail	COGS	Margin
Base (CPU-only, no GPU)	$18,500	$9,170	50%
GPU Standard (2× RTX 5090)	$24,500	$13,170	46%
GPU Pro (4× RTX 5090)	$28,500	$17,170	40%
AI Research (2× A100 80GB)	$32,000	$29,170	9%
AI Fleet (4× L40 48GB)	$42,000	$37,170	12%

> **Note:** A100/L40 configs are premium/break-even — positioned for institutional buyers. Core revenue comes from Base and GPU Standard configs.

Sales Channels

- **quillon.xyz direct** — $FORGE token burn-to-redeem (highest margin, crypto-native)
- **quillon.xyz direct purchase** — Fiat purchase option via Stripe
- **Amazon Business** — Enterprise/institutional buyers
- **Newegg Business** — Mining hardware audience
- **Crypto mining resellers** — CryptoMinerBros, MiningCave, etc.

Marketing Assets

- Whitepaper (done — `papers/quillon-forge-whitepaper.pdf`)
- Product renders (done — `docs/announcements/forge1.png`, `forge2.png`)
- Vault design language comparison (done — in whitepaper)
- Video: unboxing → power on → mining in 60 seconds
- Benchmark comparison: Forge (256-core) vs VPS (4-core) vs home PC (16-core)
- Fleet deployment guide: rack 10 units, configure via IPMI, earn at scale
- ROI calculator: hash rate × block reward × QUG price → monthly earnings

Included in Box

- 1× Quillon Forge mining machine
- 2× C14-C13 power cables (2m, 16AWG)
- 1× CAT6A Ethernet cable (3m)
- 1× USB-C to USB-A cable (management)
- 1× USB-C drive with Forge OS recovery image
- 1× Quick start card (4 steps: plug power, plug ethernet, wait 60s, done)
- 1× Rackmount rail kit (Supermicro compatible)
- 1× Serial number card with machine-ID and attestation public key
- 1× Pelican-style shipping case (reusable)

COST SUMMARY

Phase	Budget Range	Timeline
Phase 0: Design (DONE)	$0	Complete
Phase 1: System Architecture	$15–30K	Weeks 1–6
Phase 2: Mechanical Engineering	$40–80K	Weeks 4–12
Phase 3: BMC Firmware	$0–20K	Weeks 6–16
Phase 4: OS & Node Software	$0–5K	Weeks 8–16
Phase 5: First Prototype (3 units)	$25–60K	Weeks 16–20

Phase	Budget Range	Timeline
Phase 6: Burn-In + Audit + Cert	$30–80K	Weeks 20–28
Phase 7: Production (500 units)	$500K–1.8M	Weeks 28–52
TOTAL	**$610K–2.08M**	**12–18 months**

Revenue Projection (500 units @ $21,000 avg — weighted mix)

- Revenue: $10,500,000
- COGS: ~$5,500,000 (avg $11,000/unit including GPU configs)
- Gross profit: ~$5,000,000
- Breakeven on $2M investment: ~190 units
- Post-breakeven: ~$10,000 profit per additional unit sold

Revenue Projection (2,000 units @ $20,000 avg)

- Revenue: $40,000,000
- COGS: ~$16,000,000
- Gross profit: ~$24,000,000
- Net margin after R&D, marketing, support: ~40–50%

RWA TOKENIZATION — $FORGE ON-CHAIN

Token: $FORGE

- **Type**: PhysicalGoodsToken RWA
- **Backing**: Each token represents 1 physical Quillon Forge mining machine
- **Supply**: 500 tokens (matches first production batch)
- **Decimals**: 0 (indivisible — you can't own half a server)
- **Redemption**: Burn 1 $FORGE → receive configured mining machine shipped to your address
- **Contract**: Already deployed at `FORGE_TOKEN_ADDRESS` in Q-NarwhalKnight genesis

Redemption Flow

```
1. User acquires $FORGE on DEX or from Quillon


2. POST /api/v1/contracts/forge/redeem
   {
     "quantity": 1,
     "shipping_name": "...",
     "shipping_address": "...",
     "cpu_config": "epyc-9755-dual",
     "gpu_config": "rtx-5090-dual",
     "ram_gb": 512,
     "cooling_type": "liquid-copper",
     "chassis_color": "titanium-copper"
   }


3. Backend: Burns 1 FORGE from user's balance
   Creates ForgeRedemption { id: "FR-1707321600-4fff16bc" }
```

```
4. Status progression (admin updates via /forge/fulfill):
   pending → configured → assembling → testing → shipped → delivered


5. User receives Quillon Forge, powers on
   Machine auto-registers machine_id + attestation_pubkey on-chain


6. Machine begins mining QUG + serving AI inference
```

Use Cases

1. **Pre-sale funding**: Sell $FORGE tokens before production to fund manufacturing
2. **Secondary market**: Holders trade $FORGE on DEX while awaiting delivery
3. **Proof of purchase**: On-chain proof of Quillon Forge ownership
4. **Fleet management**: Token holders = verified machine operators
5. **Warranty tracking**: Machine-ID linked to token burn transaction
6. **Upgrade program**: Future models → new token series ($FORGE-V2, etc.)

Fleet Statistics API

```
GET /api/v1/contracts/forge/stats

{
  "total_supply": 500,
  "circulating": 487,
  "burned": 13,
  "redemptions": {
    "total_orders": 13,
    "total_machines_redeemed": 15,
    "pending": 3,
    "configured": 2,
    "assembling": 4,
    "testing": 2,
    "shipped": 1,
    "delivered": 3
  },
  "fleet_stats": {
    "total_cores_ordered": 3840,
    "epyc_configurations": 12,
    "xeon_configurations": 3,
    "gpu_equipped": 8
  }
}
```

APPENDIX A: CONFIGURATION OPTIONS AT REDEMPTION

Option	Choices	Price Impact
CPU	`epyc-9755-dual` (256C, default), `epyc-9654-dual` (192C, -$3K), `xeon-w9-3595x-dual` (120C, -$4K)	Base to -$4K
GPU	`none` (default), `rtx-5090-dual` (+$4K), `rtx-5090-quad` (+$8K), `a100-dual` (+$20K), `l40-quad` (+$28K)	$0 to +$28K
RAM	512GB (default), 1024GB (+$1.5K), 2048GB (+$6K)	$0 to +$6K
Storage	2×3.84TB NVMe RAID-1 (default), 2×7.68TB (+$800), 4×3.84TB RAID-10 (+$900)	$0 to +$900
NIC	100GbE (default), 10GbE (-$230)	-$230 to $0
Cooling	`liquid-copper` (default), `liquid-black` (black nickel tubing, +$100)	$0 to +$100
Chassis	`titanium-copper` (default), `titanium-black` (DLC coating, +$200), `titanium-natural` (raw Ti, -$100)	-$100 to +$200

APPENDIX B: POWER & COOLING REQUIREMENTS

Configuration	Wall Power	Heat Output	Cooling (BTU/hr)	Circuit
Base (no GPU)	1,200W	1,080W	3,685	1× 15A 120V or 1× 10A 240V
2× RTX 5090	2,350W	2,230W	7,610	1× 30A 240V
4× RTX 5090	3,500W	3,380W	11,535	2× 20A 240V

Recommended environment: Air-conditioned server room, 18–25°C ambient, <60% humidity. **Noise level:** ~50 dBA at 1m (base), ~58 dBA (GPU config, fans at max).

Document generated February 2026. QUILLON FORGE is a product of the Q-NarwhalKnight Project. https://quillon.xyz