FORSCHUNGSPAPIER I · VERSION 1.0

Model OS: Auditierbare Co-Access-Platzierung und worker-lokale GPU-Expertengraphen

Paper I · Forschungsstand auf einem bandbreitenarmen Commodity-Cluster

3 GPU-Worker	**195** Linux-Tests	**21** Messgrafiken	**26,82 %** Fresh Trace*

* EXP-006D bestätigt den unveränderten EXP-006C-Plan auf einem neuen Routertrace gegenüber der exklusiven Referenz. Keine GPU-Phase.

Verfasser: Joachim Müller-Peter
Projekt: Model OS / Project Ant Colony · itbois, Schweiz
Stand: 26. Juli 2026
Status: öffentliches Forschungspapier; nicht eingereicht; nicht peer-reviewt
Anspruchsgrenze: keine autoregressive Textgenerierung, keine verteilte Granite-Inferenz und keine Granite-Ausführung auf den Intel-GPUs

Inhalt

Der aktuelle, reproduzierbare Stand des Model-OS-Forschungsprojekts. Die Anspruchsgrenzen und negativen Ergebnisse sind Bestandteil des Manuskripts.

Abstract

Große offene Modelle überschreiten häufig den schnellen Speicher einzelner Commodity-Rechner. Eine naheliegende Reaktion ist, Gewichte, Speicher und Berechnung über mehrere Knoten zu verteilen. Auf einem alten Cluster mit 1-Gb/s-Ethernet ist jedoch nicht die nominelle Gesamtkapazität, sondern die Kombination aus Modellstruktur, Speicherresidenz, Kommunikation, Warteschlangen und Fehlergrenzen entscheidend.

Wir präsentieren Model OS, eine planungsorientierte Control Plane mit einer kleinen, auditierbaren Data Plane. Der Versuchsaufbau besteht aus acht erreichbaren Mac-mini-Proxmox-Knoten; drei 2012er Knoten führen begrenzte Intel-GPU-Workloads über gegenseitig authentifiziertes TLS 1.3 aus. Schrittweise Experimente prüfen denselben Systempfad von einem transienten GPU-Aufruf über warme Prozesse und residente Expertentensoren bis zu einem dreischichtigen, informationserhaltenden Expertengraphen.

Der akzeptierte M5.1-Graph verarbeitet zwei Micro-Batches über drei Zwei-Experten-Schichten mit 24 verifizierten Hardware-GPU-Ergebnissen. Die Pipeline reduziert die Controller-Wandzeit von 0,588918 auf 0,482790 Sekunden (18,021 Prozent), während alle 1.024 Endwerte ungleich null bleiben und serielle sowie pipelined Ausführung bitgenau übereinstimmen.

Darauf aufbauend vergleicht das präregistrierte EXP-005 drei Platzierungsregeln für acht synthetische Experten. Eine vollständige Suche über 6.561 Belegungen reduziert auf einer 10.000-Ereignis-Planungsspur die platzierungsabhängigen Remote-Schnitte um 69,5 Prozent gegenüber Round Robin und um 67,204301 Prozent gegenüber reiner Byte-Balancierung. Auf einer vorab eingefrorenen Phasenverschiebung bleiben 54,0 und 47,126437 Prozent Einsparung erhalten; bei gleichverteilten Expertenpaaren entsteht kein künstlicher Vorteil.

M6 setzt die bisher simulierte Co-Location-Voraussetzung als begrenzten Live-Pfad um. In 30 gepaarten Läufen liefert ein Worker mit zwei residenten GPU-Experten denselben Endhash wie zwei parallele Controller-RPCs. Ein lokaler RPC reduziert die medianen serialisierten JSON-Anwendungsbytes von 14.469 auf 8.098 Bytes (44,032069 Prozent), erhöht aber die mediane Controller-Wandzeit von 0,078680 auf 0,120258 Sekunden (52,843798 Prozent), weil beide Experten auf einer GPU nacheinander laufen.

EXP-006 prüft den Transfer auf einen realen Router. Der CPU-Pfad des hash-gepinnten Granite-3.1-1B-A400M-Modells erzeugt für 502 Tokens 12.048 Layer-Token-Routerereignisse und 96.384 Top-8-Auswahlen. Drei Wiederholungen stimmen bei Tokens, geordneten Experten und rohen Routerlogits exakt überein. Die darauf kalibrierte Co-Access-Platzierung reduziert Remote-Paarinzidenzen jedoch nur um 7,711811 Prozent gegenüber beiden Baselines und verfehlt damit die vorab festgelegte 10-Prozent-Schwelle. Der GPU-Replay bleibt folgerichtig gesperrt.

EXP-006B prüft als neue Präregistrierung, ob sechs vorab bezeichnete Promptdomänen diesen schwachen Effekt verdeckt haben. Zwei vollständige CPU-Durchläufe von 48 neuen Prompts ergeben exakt denselben Trace mit 1.681 Tokens und 40.344 Routerereignissen. Domänenspezifische Platzierungen sparen im ungewichteten Holdout-Mittel lediglich 0,558882 Prozent gegenüber einer global kalibrierten Co-Access-Platzierung, 6,031028 Prozent gegenüber beiden naiven Baselines und 0,875110 Prozent gegenüber einer falsch rotierten Domänenkontrolle. Alle vier vorab festgelegten Spargates scheitern.

Eine separat präregistrierte Hostreplikation wiederholt den ursprünglichen EXP-006-Vertrag auf node1. Alle Routingentscheidungen, Platzierungen, Prozentwerte und selbst die vollständigen float32-Routerlogits stimmen bitgenau mit node3 überein. Damit ist der negative 7,711811-Prozent-Befund hostreproduzierbar. Seine Aussage bleibt jedoch auf exklusive Platzierung mit genau einem Heimat-Node je Experte begrenzt; überlappende Expertengruppen mit gezielten Replikaten wurden nicht untersucht.

EXP-006C untersucht genau diesen neuen Planungsraum offline. Ein vor Code festgelegter exakter Planner verteilt zwölf Slots als vier überlappende Dreiergruppen und repliziert vier der acht ausgewählten Experten. Gegenüber dem besten exklusiven 3-Node-Plan sinken Remote-Paarinzidenzen in der Kalibrierung von 13.104 auf 9.260 (29,334554 Prozent) und im gesperrten Holdout von 9.069 auf 6.576 (27,489249 Prozent). Kapazitätsgleiche replizierte Kontrollen, Uniformspur, Laufzeit und Fail-closed-Ausfallklassifikation bestehen ebenfalls. Vier Nodes ohne Zusatzreplikate sind dagegen schlechter als die 3-Node-Referenz.

EXP-006D prüft denselben Plan ohne Neuoptimierung auf einem vor dem Modelllauf eingefrorenen neuen Korpus. Zwei vollständige CPU-Durchläufe von 48 Prompts ergeben exakt denselben Trace mit 1.633 Tokens, 39.192 Routerereignissen und 313.536 Top-8-Auswahlen. Der feste Kandidat senkt 62.446 Remote- Paarinzidenzen der

exklusiven Referenz auf 45.697, also um 26,821574 Prozent. Er spart 23,612992 Prozent gegen beide kapazitätsgleichen Round-Robin- und Byte-/Frequenzkontrollen und 24,189588 Prozent gegen die Niedrigfrequenzkontrolle. Alle acht Domänen sind besser; die untere 95-Prozent-Grenze eines vorab festgelegten 10.000-fachen Prompt-Bootstraps liegt bei 26,191329 Prozent. Alle Bestätigungsgates bestehen.

EXP-006E sollte diesen positiven Offline-Plan in einen begrenzten authentifizierten Vier-Failure-Domain-GPU-Canary übertragen. Der letzte read-only Preflight meldete jedoch für den geschützten Voice-Container CT333 `pressurecpufull=0.28`, während die präregistrierte Grenze exakt null verlangte. Das Experiment endete deshalb vor der offiziellen Node8-Qualifikation als `node8_ineligible_before_live_execution`. Es wurden null Replays und null GPU-Expertenjobs ausgeführt; die 7,5-Prozent-Bytefrage blieb ungemessen. Der Befund belegt das Greifen der Schutzgrenze, nicht die Data-Plane-Wirkung des Vier-Node-Plans.

EXP-006 belegt vollständige, nicht-generierende CPU-Prompt-Forwards des Granite-MoE einschließlich ausgewählter Expert-MLPs; für das Experiment wurden nur die Routerlogits ausgewertet. Nicht belegt sind autoregressive Textgenerierung, verteilte Granite-Inferenz oder Granite-Ausführung auf den Intel-GPUs. Die Ergebnisse zeigen eine reproduzierbare Systemgrenze: Eine modellbewusste Control Plane kann auf knapper Commodity-Hardware niedrigere Kommunikationsschnitte finden, und eine begrenzte GPU-Data-Plane kann den zugehörigen synthetischen Expertengraphen korrekt ausführen und lokale Co-Access-Paare mit weniger Anwendungsdaten bedienen. Der große synthetische Platzierungseffekt überträgt sich unter exklusiver Platzierung jedoch nicht in der präregistrierten Stärke auf den realen Routertrace. Gezielte Replikate zeigen dort einen positiven Offline-Effekt, den EXP-006D auf einem neuen, output-unbekannten Trace bestätigt. EXP-006E materialisiert diesen Effekt nicht: Das Schutzgate sperrt den Live-Pfad vor der Metrik. Verteilte Granite-MoE-Inferenz, reale Granite-Experten auf den Intel-GPUs und eine Wire-Byte-Messung bleiben notwendig; M6 zeigt ausdrücklich keinen Latenzgewinn.

1. Einleitung

Die übliche Beschreibung verteilter Modellinferenz beginnt mit Rechenleistung: Wie viele FLOPS, wie viel VRAM und welche Beschleunigerverbindung stehen zur Verfügung? Für einen Cluster alter Commodity-Knoten greift diese Perspektive zu kurz. Dort können lokale SSDs zusammen große Gewichtsbestände aufnehmen, während RAM, GPU-Speicher und 1-Gb/s-Ethernet eine interaktive Ausführung weiterhin ausschließen.

Model OS behandelt deshalb ein Modell nicht als große Datei, sondern als versionierten Ausführungsgraphen. Experten, dichte Stufen, Gewichts-Shards, Aktivierungen, Caches und Puffer werden zu Ressourcen mit messbaren Kosten und expliziten Lebenszyklen. Die Control Plane plant langsam und erklärbar; die Data Plane führt nur vorab zugelassene, streng begrenzte Arbeit aus.

Das vorliegende Papier untersucht sechs engere Fragen:

1. Kann eine kleine verteilte GPU-Data-Plane auf alter Hardware einen mehrschichtigen Expertengraphen korrekt, reproduzierbar und parallel ausführen?
2. Kann eine modellbewusste Offline-Platzierung auf einem eingefrorenen Co-Access-Graphen weniger potenziell vermeidbare Remote-Kommunikation erzeugen als Round Robin und Byte-Balancierung?
3. Kann worker-lokale Co-Access-Ausführung diese Kommunikationsreduktion für einen begrenzten Live-Job materialisieren, ohne die numerische Identität zu verlieren?
4. Bleibt der Platzierungsvorteil auf einem deterministisch erfassten Routertrace eines realen offenen MoE groß genug, um einen nachgelagerten GPU-Replay zu rechtfertigen?
5. Können begrenzte Replikate und überlappende Expertengruppen den schwachen exklusiven Real-Trace-Befund verbessern, ohne einen Vorteil aus einer schlechteren Mehr-Node-Baseline vorzutäuschen?
6. Kann der bestätigte Offline-Plan unter den vorab festgelegten Schutzgrenzen überhaupt zu einem Vier-Failure-Domain-Live-Replay zugelassen werden, bevor seine Data-Plane-Wirkung gemessen wird?

Der Beitrag liegt nicht in der Behauptung, alte Intel-GPUs ersetzten moderne Beschleuniger. Er liegt in der auditierbaren Verbindung von Messung, Platzierung, Sicherheitsgrenze, negativem Ergebnis und reproduzierbarer Korrektur.

2. Beitrag und Anspruchsgrenze

Der aktuelle Beitrag umfasst:

- ein versioniertes Hardware- und Kostenprofil eines heterogenen Mac-mini-Clusters;
- eine GPU-only Data Plane mit festem Kernelvertrag, mTLS 1.3, HMAC-/Replay-Schutz, begrenzter Warteschlange und ohne CPU-Rechenfallback;
- einen residenten Expertentensor-Pfad mit unabhängiger Controller-Referenzprüfung;
- einen explizit gepufferten, dreischichtigen Expertengraphen einschließlich dokumentiertem Fehlschlag und separat präregistrierter Korrektur;
- einen deterministischen, vollständigen Co-Access-Platzierungsvergleich mit Holdout und Kapazitätssensitivität;
- einen live gemessenen worker-lokalen Zwei-Experten-Pfad mit bitgleicher Remote-Baseline, unabhängigen Referenzen und explizitem Byte-/Latenzkontrast;
- einen hash-gepinnten realen Granite-Routertrace mit drei bitgenauen Wiederholungen, Kalibrierungs-/Holdout-Trennung und einem erhaltenen präregistrierten Negativergebnis;
- einen präregistrierten exakten replica-aware Planner mit absoluten und kapazitätsgleichen Kontrollen, vor dem Holdout festgeschriebenem Plan sowie Uniform-, Speicher- und Fail-closed-Ausfallprüfung;
- eine Bestätigung des unveränderten Plans auf einem neuen, vor Modelloutput eingefrorenen Routertrace mit promptbasiertem Bootstrap;
- einen separat präregistrierten Vier-Failure-Domain-Transfervertrag, dessen Schutzgate vor der ersten offiziellen Qualifikation korrekt fail-closed abbricht und die Bytefrage ausdrücklich ungemessen lässt;
- ein reproduzierbares Evidenz- und Abbildungssystem für spätere Veröffentlichung.

Nicht beansprucht werden:

- autoregressive Textgenerierung oder produktive End-to-End-Bereitstellung;
- verteilte Granite-Inferenz oder Attention-, KV-Cache- und Expert-MLP-Ausführung auf den Intel-GPUs;
- Ausführung realer Granite-Expertengewichte auf den Intel-GPUs;
- gemessene End-to-End-Inferenz- oder Latenzverbesserung durch EXP-005/M6;
- allgemeine Überlegenheit alter Hardware;
- Live-Replikation, allgemeines N+1-Failover oder produktive Modellbereitstellung;
- ein gemessener Vier-Failure-Domain-Data-Plane- oder Bytevorteil aus EXP-006E;
- Neuheit von Auto-Tuning, gelerntem Scheduling oder Kommunikationsreduktion an sich.

3. Verwandte Arbeiten und Neuheitsgrenze

Hardwarebewusste Compiler wie TVM und Ansor etablieren lern- und suchgestützte Optimierung von Tensorprogrammen (Chen et al., 2018; Zheng et al., 2020). Gandiva und Decima zeigen laufzeitbewusste beziehungsweise gelernte Clusterplanung (Xiao et al., 2018; Mao et al., 2019). Local SGD und Deep Gradient Compression reduzieren Kommunikationshäufigkeit oder -volumen (Stich, 2019; Lin et al., 2018). AlphaTensor, AlphaDev und AutoML-Zero belegen, dass maschinelle Suche Algorithmen oder Primitive entdecken kann (Fawzi et al., 2022; Mankowitz et al., 2023; Real et al., 2020).

Model OS beansprucht keine dieser Komponenten als neu. Die Kandidatenlücke liegt in ihrer begrenzten Verbindung: Modellgraph und Live-Ressourcen werden auf realer, älterer und bandbreitenbegrenzter Hardware in einen deklarativen Plan übersetzt; die Data Plane akzeptiert nur maschinell prüfbare Arbeit; jede Leistungsbehauptung bleibt an eine eingefrorene Konfiguration und negative Gates gebunden.

Der Literaturteil ist für eine Einreichung noch unvollständig. Insbesondere müssen aktuelle Arbeiten zu MoE-Expert-Placement, Expert Parallelism, Hierarchical All-to-All, Edge Inference und Co-Location systematisch ergänzt werden. Bis dahin wird die Formulierung „Kandidatenlücke“ verwendet, nicht „erstmals“.

4. Systemmodell

4.1 Control Plane

Die Control Plane erfasst:

- Online-Zustand und geschützte Produktionslasten;
- policy-sicheren RAM statt bloß physisch eingebauten RAM;
- SSD-Klassen, Linktopologie und gemessene Parallelität;
- zugelassene GPU-Backends und deren exakten Kernelumfang;
- Modellstruktur, Expertengruppen und Pufferverträge;
- Netzwerk-, Queue-, Wärme- und Fehlerkosten.

Sie erzeugt einen unveränderlichen Placement-Plan oder einen präzisen Unmöglichkeitsbefund. Ein Plan wird nicht allein dadurch interaktiv, dass seine Gewichte auf SSD passen.

4.2 Data Plane

Die produktive M3-Grenze akzeptiert nur einen hash-gepinnten Matrixmultiplikations-Canary. Ein nicht produktiver M4/M5-Pfad erweitert dies um einen unveränderlichen 16×16-Expertentensor und begrenzte Aktivierungen. Jeder Worker:

- läuft unprivilegiert als `modelos-gpu`;
- sieht ausschließlich das Render-Gerät;
- besitzt 384 MiB RAM-, 75-Prozent-CPU- und 32-Task-Grenzen;
- führt höchstens einen GPU-Job gleichzeitig aus;
- akzeptiert weder Quellcode noch Kommandofelder;
- verwirft Software-Renderer und unbekannte Kernel;
- beweist GPU-Ausführung und lässt das Ergebnis unabhängig prüfen.

M6 erweitert diesen Laborpfad auf einem getrennten Canary-Port um genau zwei unveränderliche residente Experten je Worker. Die Expertenarithmetik bleibt GPU-pflichtig; nur der kleine festgelegte Residual-Merge darf auf der Worker-CPU laufen. M6 akzeptiert weder beliebige Gewichte noch dynamische Expertenlisten.

EXP-006 ergänzt keine produktive Data Plane. Ein isolierter, netzloser Referenzlauf auf node3 führt das gepinnte Granite-Modell auf der CPU als vollständigen Prompt-Forward aus, einschließlich Attention, Router und ausgewählten Expert-MLPs. Nur die Routerlogits werden ausgewertet; autoregressive Generation findet nicht statt. Ein separater GPU-Canary dürfte erst nach bestandenem Platzierungsgate starten; dieses Gate wurde nicht erreicht.

EXP-006B verwendet dieselbe isolierte CPU-Ausführungsgrenze, dasselbe Modell und dieselbe Runtime, aber einen vollständig neuen 48-Prompt-Korpus. Das Experiment bleibt auch bei Erfolg auf Routertrace und Offline-Platzierung begrenzt; es autorisiert weder GPU-Replay noch Live-Deployment.

EXP-006C führt keinen neuen Modelllauf aus. Es verwendet die eingefrorene EXP-006-Paarprojektion und trennt die Auswertung in zwei Programmeingänge. Der erste sieht nur Kalibrierung, durchsucht exakt kapazitätsgültige Expert-zu-Node-Mengen und schreibt einen unveränderlichen Plan-Hash. Erst der zweite Eingang öffnet Holdout, Uniformspur und Ausfälle. Der Hauptfall umfasst vier logische Nodes, zwölf Slots und höchstens zwei Kopien je Experte. Die logische Verwendung von node4 ist keine M3- oder Granite-GPU-Qualifikation.

EXP-006D verwendet wieder ausschließlich den isolierten CPU-Capture. Der EXP-006C-Plan und alle Kontrollen stehen vor dem neuen Modelloutput fest; der Capture darf weder Experten neu wählen noch das Placement optimieren.

EXP-006E definiert getrennte, nicht produktive Canary-Ports für eine exklusive Drei-Node-Referenz und den festen replizierten Vier-Node-Plan. Der vierte logische Slot wird physisch auf node8 abgebildet. Bevor irgendein Worker startet, müssen statische Host-, GPU-, Voice-, Swap-, Temperatur- und Pressure-Gates bestehen. Das CPU-Full-Pressure-Gate von CT333 besteht nicht; darum entsteht in EXP-006E keine laufende Data Plane und kein Live-Datensatz.

4.3 Puffer- und Parallelitätsmodell

Der Mehrschicht-Controller verwendet zwei Aktivierungspuffer mit explizitem Lebenszyklus:

`FREE` → `LOADING` → `READY` → `IN_USE` → `RELEASABLE` → `FREE`.

CPU-Vorbereitung, Netzwerktransport und GPU-Arbeit dürfen sich überlappen, solange keine Komponente einen GPU-eigenen Puffer überschreibt. Backpressure verhindert unbeschränkte Queue- oder RAM-Ausweitung.

5. Experimentierplattform

Der Cluster umfasst acht erreichbare Proxmox-Knoten im privaten 1-Gb/s-Netz. Für den aktuellen GPU-Pfad sind node1, node2 und der thermisch begrenzte node3 zugelassen. Alle drei verwenden integrierte Intel HD Graphics

1 Node3 arbeitet wegen eines ungeklärten Wärmeproblems ohne CPU-Turbo und mit Routing-Penalty; unbegrenzte Laufzeiten sind nicht zugelassen.

Node8 besitzt eine Intel Iris Graphics 5100 und bestand zuvor einen isolierten Hardware-GPU-Canary. Für EXP-006E trägt er nur den vorgesehenen logischen node4-Slot. Das qualifiziert ihn nicht allgemein als Model-OS-Worker. CT333 trägt gleichzeitig geschützte Voice-Dienste; deren Zustand hat Vorrang vor einer Forschungsmetrik.

Die Plattform ist absichtlich kein Leistungsersatz für moderne Beschleuniger. Sie dient als stressreicher Prüfstand für:

- knappen Resident-Speicher;
- hohe Kommunikationskosten;
- heterogene Hintergrundlast;
- thermische Grenzen;
- reale SSD- und RAM-Parallelität;
- reproduzierbare Fehler- und Abbruchpfade.

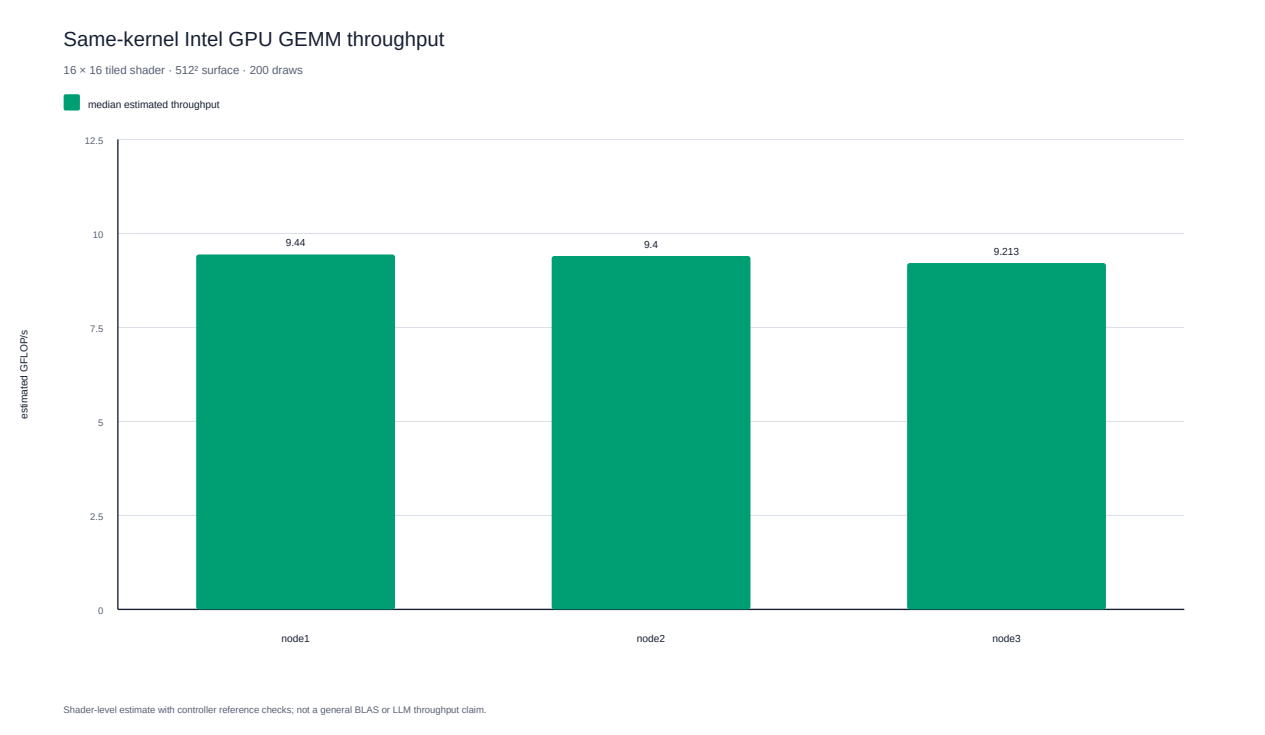

Abbildung 1. Geschätzter Durchsatz des identischen Intel-GPU-GEMM-Kernels.

Node3 liegt beim identischen GEMM-Canary trotz CPU-Drosselung nur 2,412 Prozent unter node1.

6. Methodik

6.1 Plan-first und Präregistrierung

Jeder Meilenstein trennt:

1. Hypothese und Akzeptanzgrenze;
2. Implementierung und Vertragstests;
3. offiziellen Lauf;
4. unveränderliche Evidenz;
5. Interpretation und nächste Grenze.

Fehlgeschlagene Läufe werden nicht gelöscht. M4 bewahrte einen Speicherwachstumsfehler, M5 einen numerischen Kollaps und EXP-005 eine zunächst falsche Test-Fixture-Erwartung. EXP-006 bewahrt sowohl einen abgebrochenen Paketversuch mit unbenötigten CUDA-Abhängigkeiten als auch das verfehlte 10-Prozent-Platzierungsgate. EXP-006E bewahrt drei präqualifikatorische Startdiagnosen, den bestandenen Implementierungsvertrag und den anschließenden fail-closed Preflight-Ausgang, ohne daraus einen Live-Messwert zu konstruieren.

6.2 Metriken

Je nach Meilenstein werden gemessen:

- Controller-Wandzeit und Worker-Roundtrip;
- GPU-Engine- und Queue-Zeit;
- Aktivierungs-, JSON- und gesamte Anwendungsbytes;
- Speicherwachstum, Swap und Temperatur;
- CPU-, I/O- und Memory-Full-Pressure geschützter Gäste;
- GPU-, mTLS-, HMAC-, Expert-, Gewichts- und Referenznachweise;
- Remote-Schnitte, Nachrichten, maximale Belegung und Planungszeit;
- Routerereignisse, geordnete Top-k-Experten, rohe Logit-Hashes und Top-8-zu-Rang-9-Abstände.

6.3 Baselines

EXP-005 vergleicht:

- Round Robin;
- Largest-first Byte Balancing;
- exhaustive Co-Access-Cost-Platzierung.

Alle Strategien erhalten dieselben acht Experten, dieselben drei Nodes und dieselben Slot- und Byte-Grenzen. Der Co-Access-Optimierer sieht nur die Planungsspur; Holdout und Uniformspur werden ausschließlich ausgewertet.

M6 vergleicht für 30 vorab festgelegte Aktivierungen zwei Rechenpfade:

- zwei parallele mTLS/HMAC-RPCs zu je einem residenten Experten auf node1 und node2 mit Merge im Controller;
- einen RPC zu zwei residenten Experten auf node1 mit serieller GPU-Ausführung und identischem Merge auf der Worker-CPU.

Die Reihenfolge alterniert je Seed. Primärmetrik sind die serialisierten JSON-Request- und Response-Körper; Controller-Wandzeit ist deskriptiv.

EXP-006 verwendet das öffentliche `ibm-granite/granite-3.1-1b-a400m-base` an Revision `408b6e90baab8cf24f4aa9f8e19703ffa0a53b29`. Ein vorab eingefrorener 16-Prompt-Korpus wird in Kalibrierung und Holdout geteilt. Drei CPU-Läufe müssen Token-IDs, geordnete Top-8-Experten und rohe float32-Routerlogits exakt wiederholen. Nur die Kalibrierung bestimmt die acht häufigsten Experten und die exhaustive Co-Access-Platzierung; Holdout und Uniformkontrolle werden nur ausgewertet. Ein GPU-Replay ist nur zulässig, wenn Co-Access Cost in der Kalibrierung mindestens 10 Prozent gegenüber Round Robin und Byte Balancing spart, im Holdout nicht schlechter ist und die Uniformkontrolle nicht verschlechtert.

EXP-006C erweitert die Platzierung um begrenzte Replikate. Der Hauptfall verteilt zwölf Slots über vier logische Nodes, erlaubt höchstens zwei Kopien je Experte und durchsucht 745.920 gültige Belegungen exakt. Eine absolute exklusive 3-Node-Referenz und kapazitätsgleiche replizierte Kontrollen verhindern, dass zusätzliche Nodes oder Slots allein als Placement-Erfolg gezählt werden.

EXP-006D wendet den vor dem Modelllauf festgeschriebenen EXP-006C-Plan ohne Neusuche auf 48 neue Prompts aus acht Domänen an. Primärmetrik ist die Reduktion der Remote-Paarinzidenzen gegenüber der exklusiven Referenz. Zusätzlich werden alle Domänen und ein deterministischer 10.000-facher Prompt-Bootstrap ausgewertet.

EXP-006E bindet den EXP-006D-Paartrace an einen festen 112-Fall-Replay mit acht verschiedenen synthetischen GPU-Experten. Lokale Paare verwenden einen, Remote-Paare zwei authentifizierte JSON-RPCs. Primärmetrik wären die mit den eingefrorenen Paarhäufigkeiten gewichteten Request- und Response-Body-Bytes; gefordert sind mindestens 7,5 Prozent Einsparung. Vor jeder Metrik stehen jedoch harte Node8- und CT333-Schutzgates. Ein einziger Gatefehler beendet das Experiment vor Qualifikation und Replay.

7. Ergebnisse

7.1 Von transientem Dispatch zur geschützten warmen Data Plane

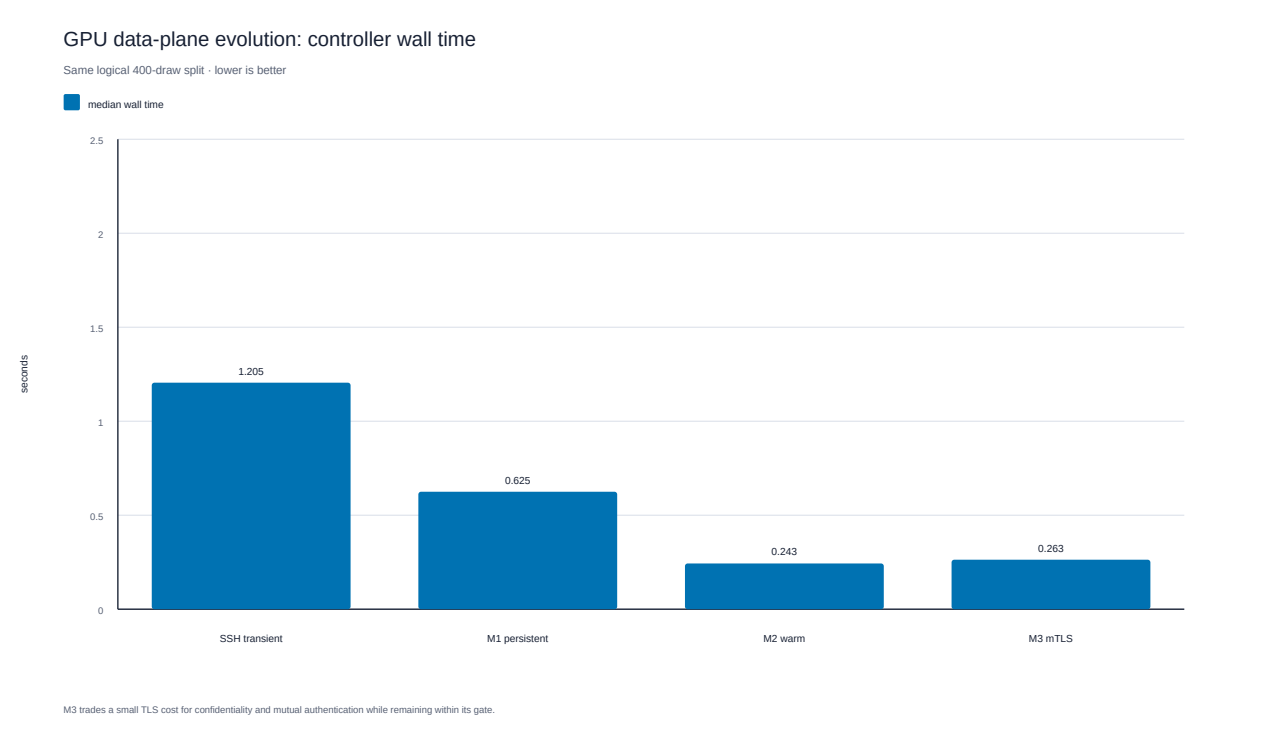

Abbildung 2. Entwicklung der Controller-Wandzeit vom transienten Aufruf bis mTLS.

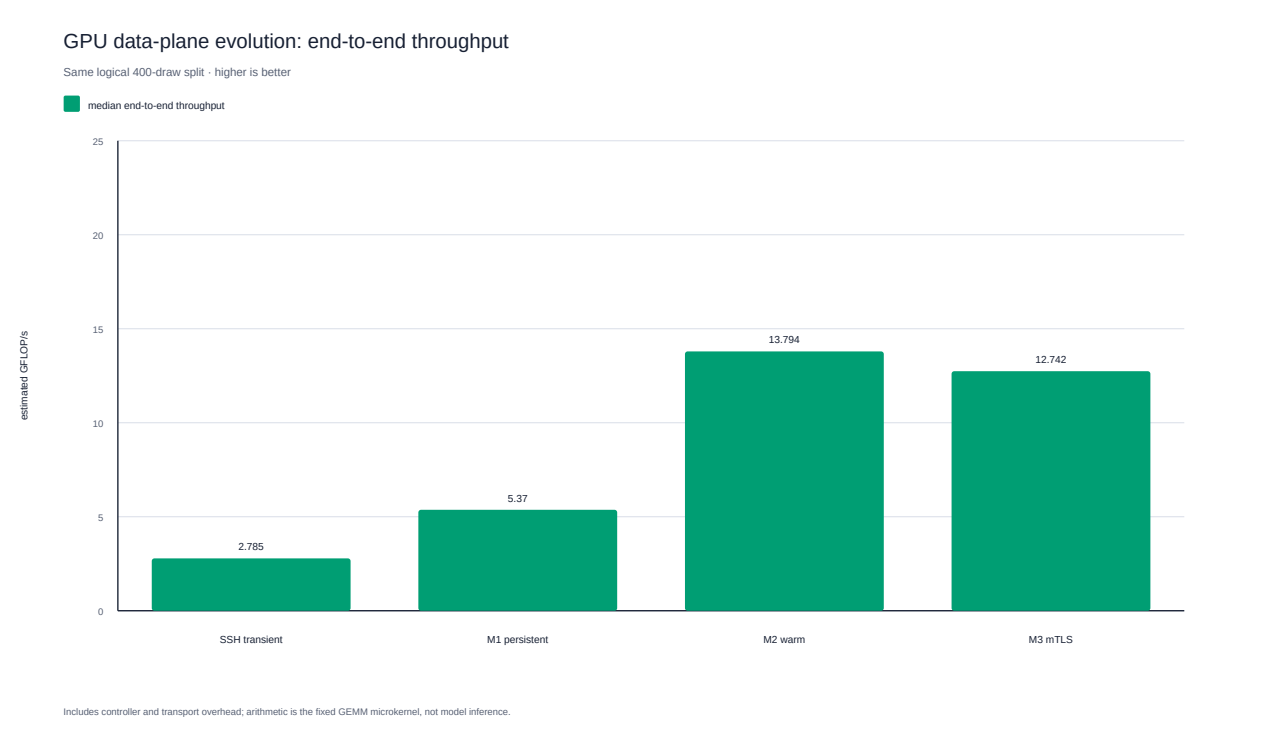

Abbildung 3. Entwicklung des geschätzten End-to-End-Durchsatzes der GPU-Data-Plane.

Der Übergang vom transienten SSH-Aufruf zu M1 reduziert die Median-Wandzeit des gleichen logischen 400-Draw-Splits von 1,204981 auf 0,624857 Sekunden. Der warme M2-Pfad erreicht 0,243252 Sekunden. M3 ergänzt mTLS 1.3 und benötigt 0,263336 Sekunden; der kleine Rückgang gegenüber M2 ist der Preis der Vertraulichkeits- und Identitätsgrenze, nicht ein verdeckter Baselinewechsel.

7.2 Speicher- und I/O-Parallelität

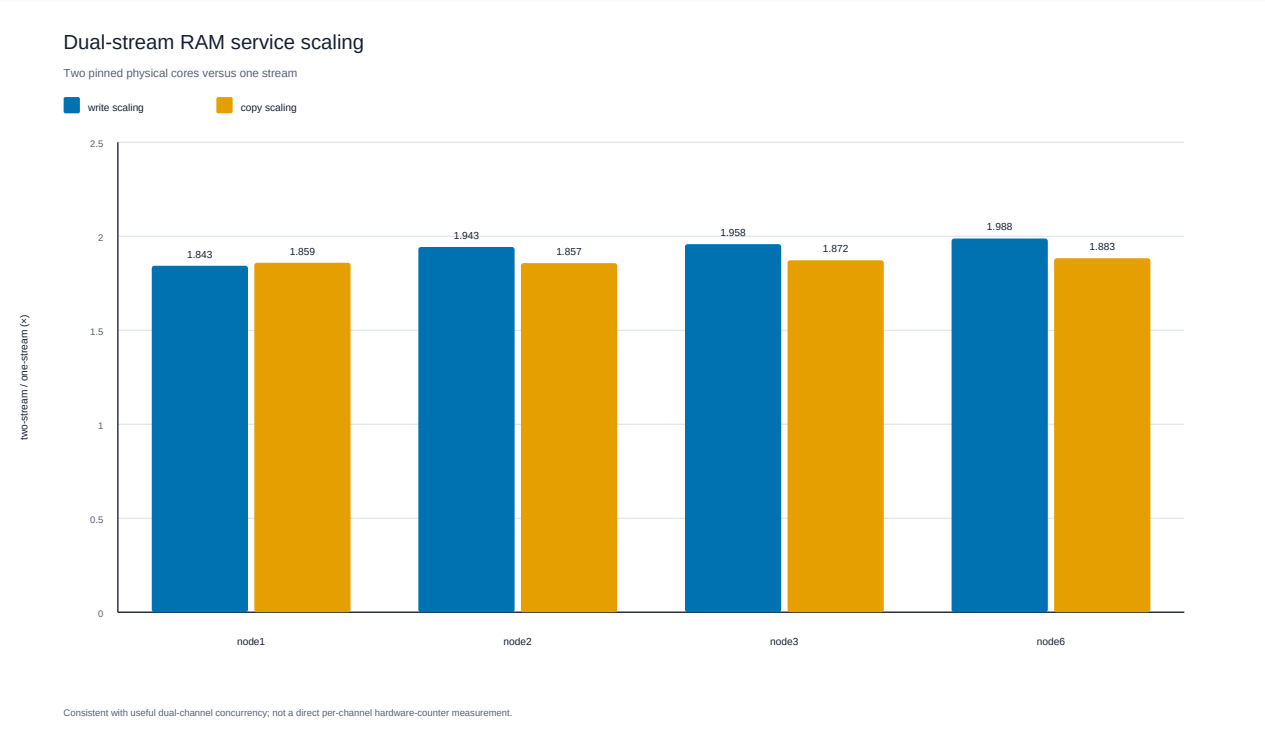

Abbildung 4. Skalierung gleichzeitiger RAM-Streams auf vier 2012er Quad-Core-Nodes.

Vier 2012er Quad-Core-Nodes erreichen beim Schreiben eine Zwei-Stream-Skalierung von 1,843× bis 1,988× und beim Kopieren von 1,857× bis 1,883×. Dies belegt nützliche gleichzeitige RAM-Bedienung, nicht direkt zwei separat gemessene physische Kanäle.

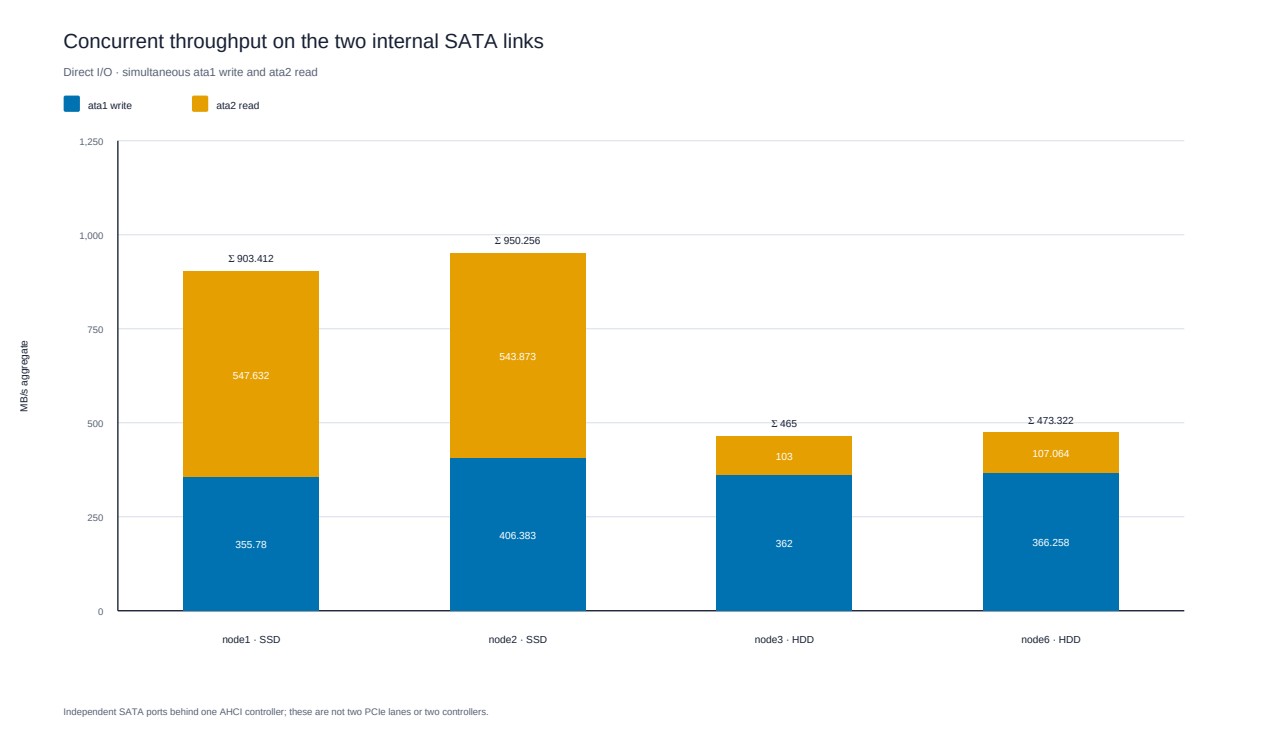

Abbildung 5. Gleichzeitiger direkter Durchsatz der beiden internen SATA-Links.

Node1 und node2 erreichen mit zwei SSDs 903 beziehungsweise 950 MB/s gleichzeitigen direkten Write-/Read-Durchsatz. Node3 und node6 kombinieren SSD und HDD und erreichen 465 beziehungsweise 473 MB/s. Die beiden Ports hängen am selben AHCI-Controller und dürfen nicht als zwei PCIe-Lanes beschrieben werden.

7.3 Residente Experten und Fan-out

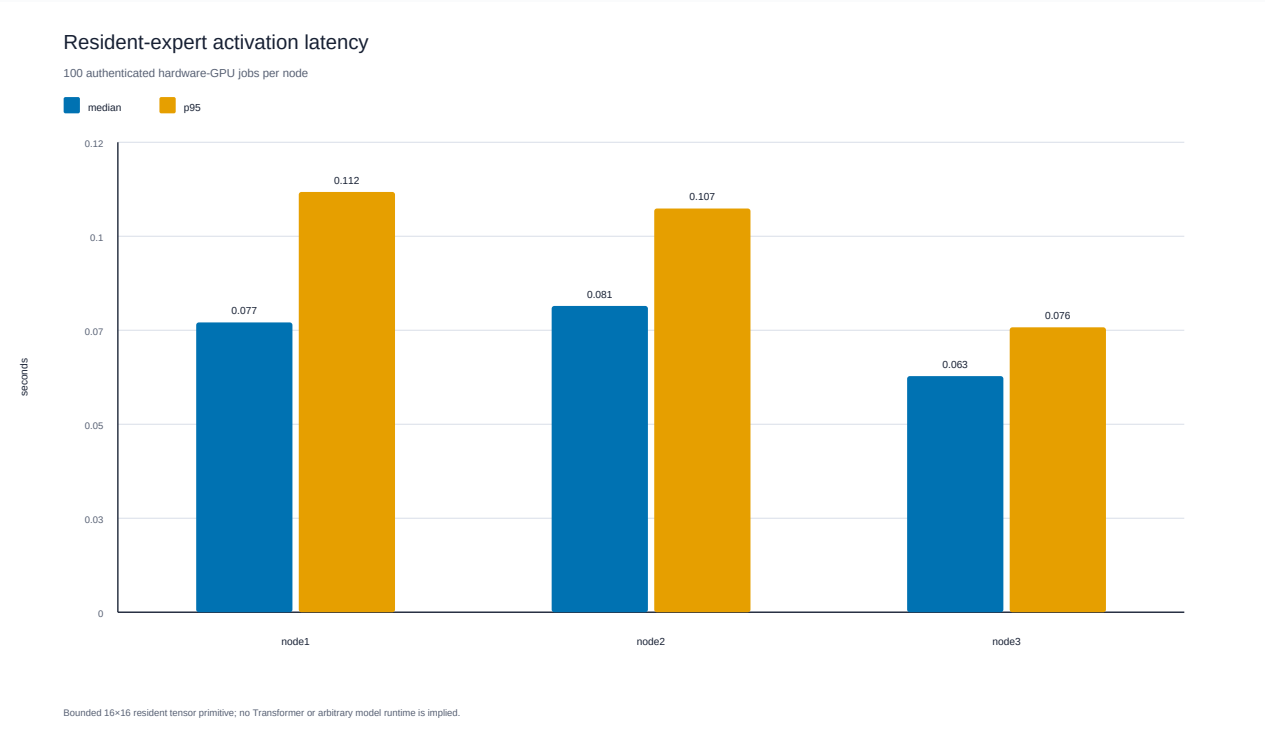

Abbildung 6. Latenz der residenten Expertenaktivierung auf den drei GPU-Workern.

Jeder der drei Worker besteht 100 Aktivierungsjobs mit genau einem Gewichts-Upload und stabiler Ressourcenidentität. Die Median-Roundtrips liegen zwischen 0,062801 und 0,081465 Sekunden.

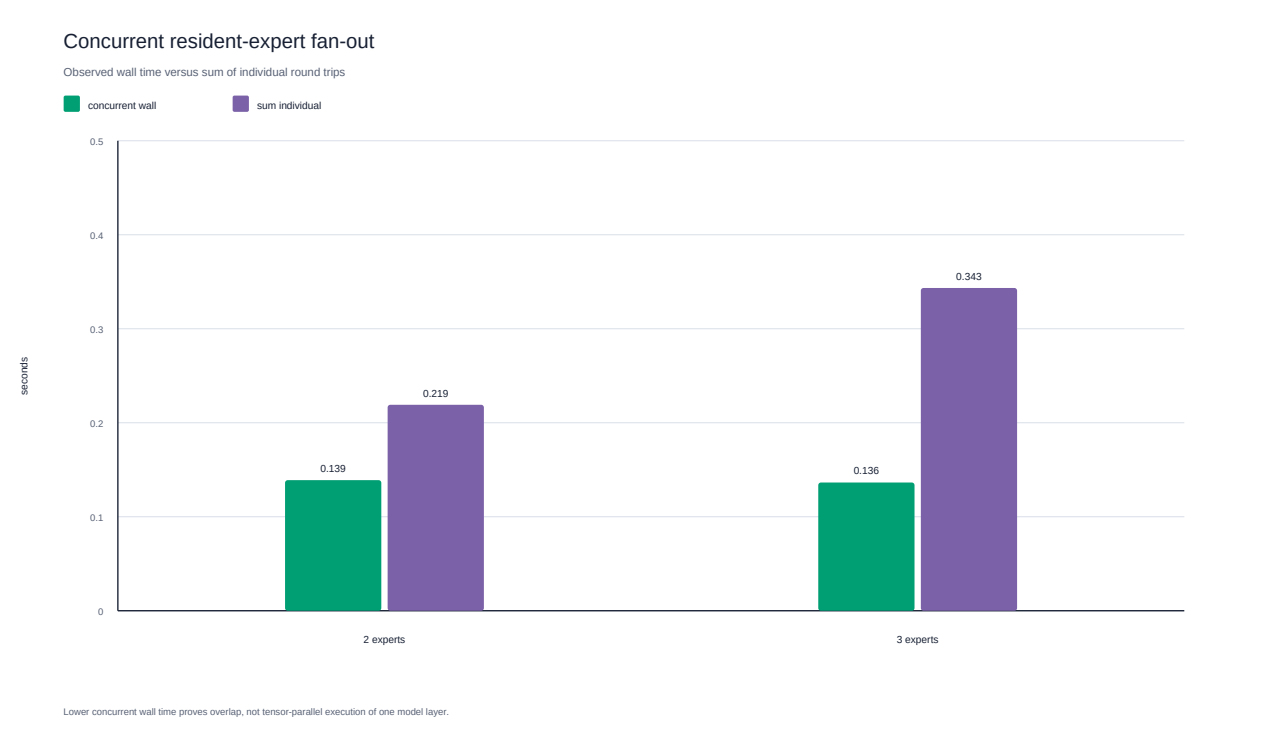

Abbildung 7. Überlappung des gleichzeitigen Fan-outs auf zwei und drei residente Experten.

Der Zwei-Experten-Fan-out benötigt 0,138936 Sekunden statt der Summe von 0,219112 Sekunden. Der Drei-Experten-Fan-out benötigt 0,136409 statt 0,343379 Sekunden. Das beweist Überlappung unabhängiger Expert-Aufrufe, nicht Tensorparallelität innerhalb eines Operators.

7.4 M5-Fehlschlag und M5.1-Korrektur

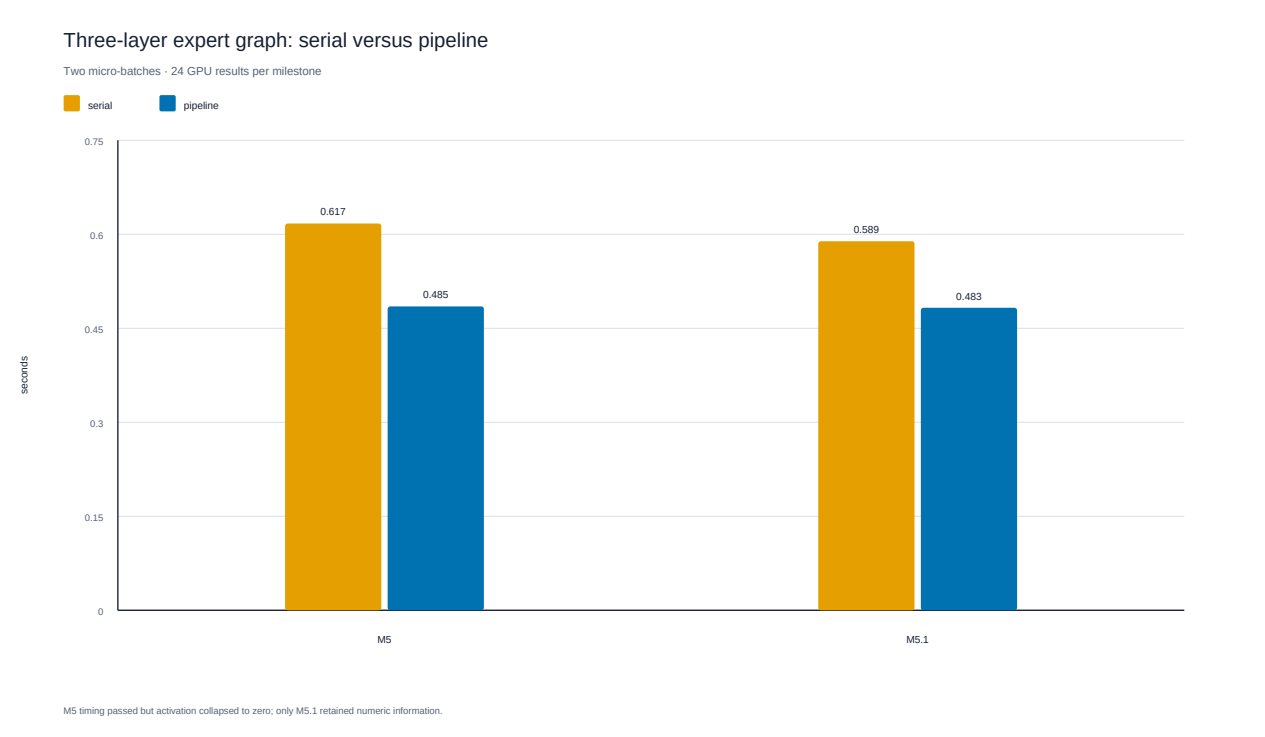

Abbildung 8. Wandzeit des seriellen und gepipelten dreischichtigen Expertengraphen.

M5 besteht Transport-, Identitäts-, Puffer- und Timing-Gates, verliert aber nach Schicht 1 sämtliche numerische Information. Ursache ist eine zweite 31/255-Skalierung auf bereits byte-kodierten GPU-Ausgaben.

M5.1 wird separat präregistriert. Der Residual-Merge

```
min(31, floor((2·input + expert1 + expert2 + 2) / 4))
```

bewahrt in beiden Micro-Batches alle 1.024 Endwerte. Der Pipeline-Pfad ist 18,021 Prozent schneller als der serielle Pfad und liefert denselben Endhash.

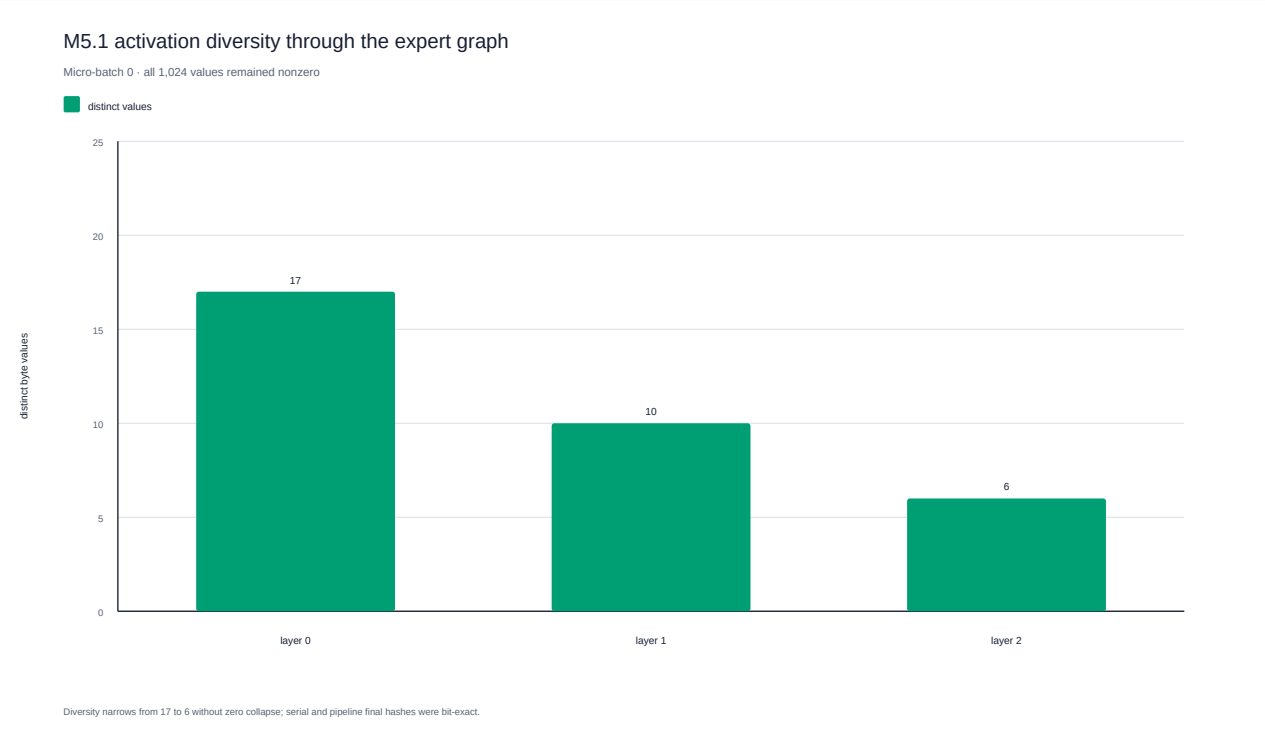

Abbildung 9. Wertevielfalt der M5.1-Aktivierungen über die drei Expertenschichten.

Die Zahl verschiedener Werte sinkt kontrolliert von 17 über 10 auf 6; sie kollabiert nicht auf null.

7.5 EXP-005 Platzierung

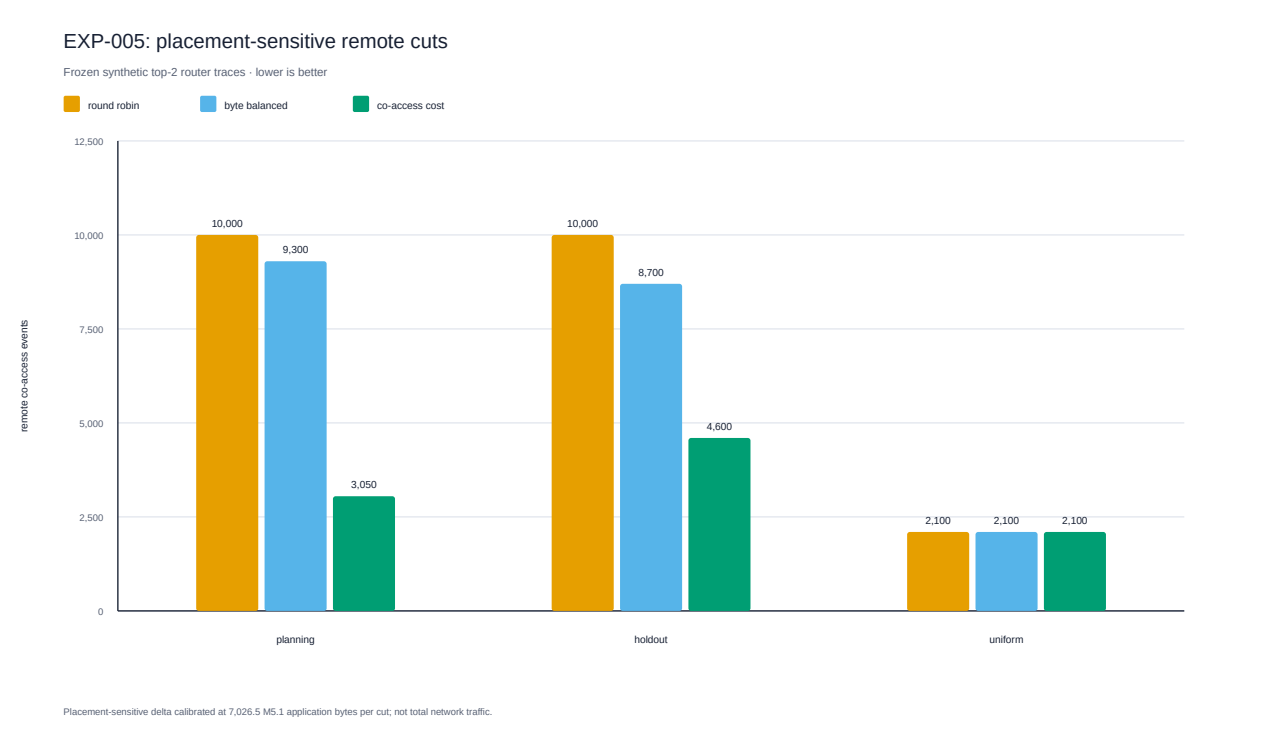

Abbildung 10. Platzierungsabhängige Remote-Schnitte auf Planung, Holdout und Uniformspur.

Auf der Planungsspur erzeugen Round Robin, Byte Balancing und Co-Access Cost 10.000, 9.300 und 3.050 Schnitte. Mit 7.026,5 gemessenen M5.1- Anwendungsbytes pro potenziell vermeidbarem Austausch entspricht dies

einer platzierungsabhängigen Differenz von 48.834.175 Bytes gegenüber Round Robin.

Auf dem Holdout bleiben 4.600 statt 10.000 beziehungsweise 8.700 Schnitte. Die Uniformspur endet für alle Strategien bei 2.100; die Optimierung gewinnt dort also nicht durch einen asymmetrischen Vergleich.

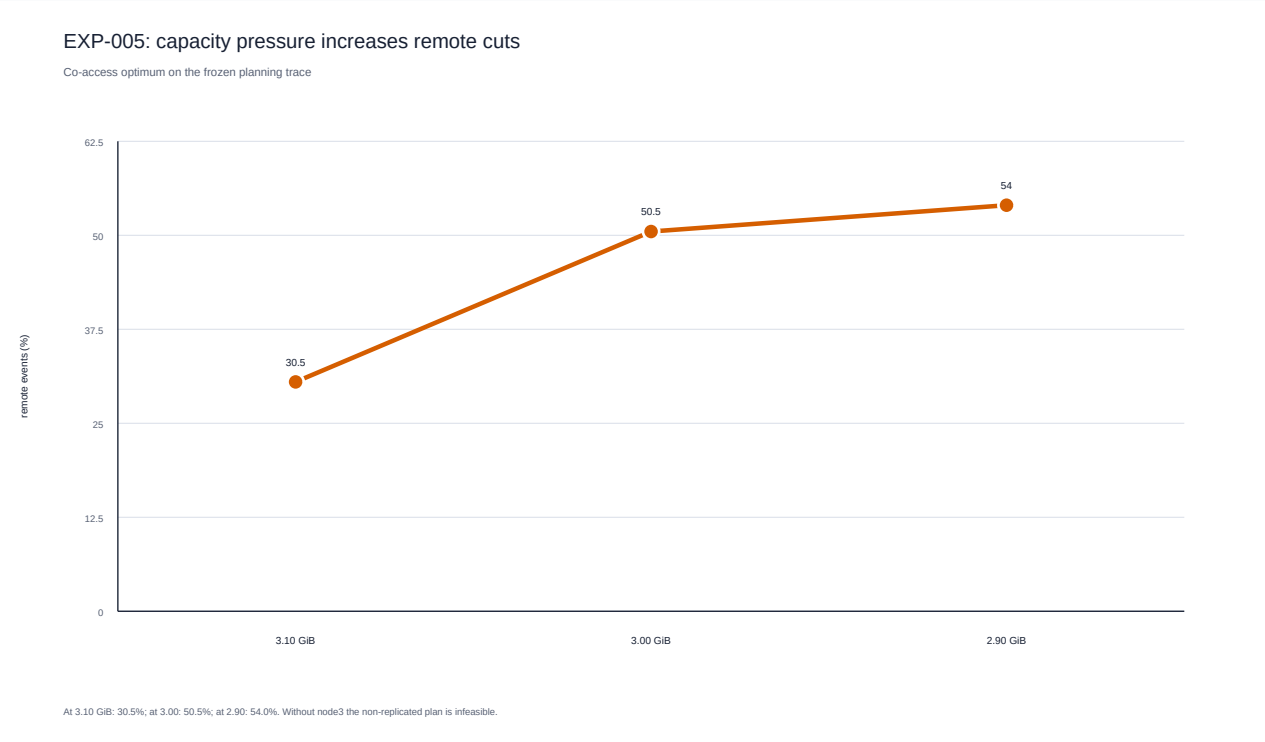

Abbildung 11. Remote-Schnitte von EXP-005 unter zunehmendem Kapazitätsdruck.

Schon 0,10 GiB weniger Expert-Kapazität pro Node erhöhen den optimalen Remote-Anteil von 30,5 auf 50,5 Prozent. Bei 2,90 GiB steigt er auf 54,0 Prozent. Ohne node3 ist der Plan wegen sechs verbleibender Slots für acht Experten unzulässig.

7.6 M6 Live-Co-Access

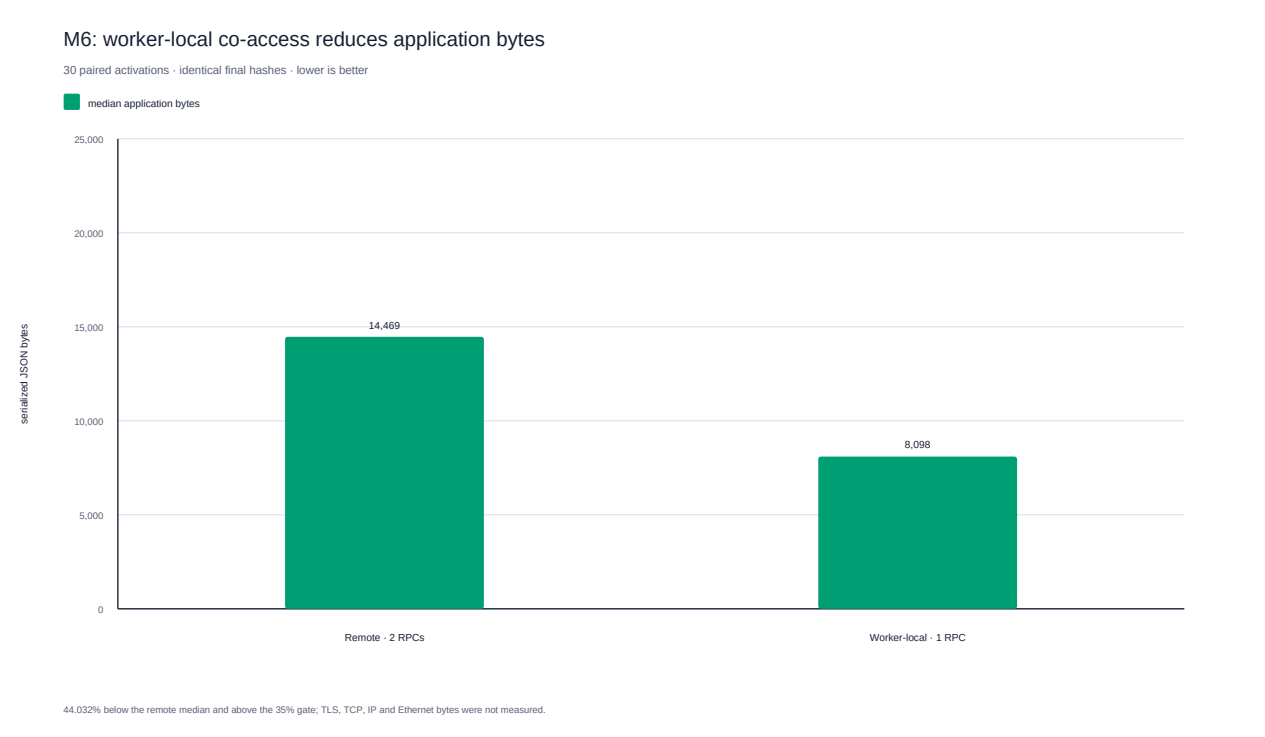

Abbildung 12. Serialisierte JSON-Anwendungsbytes der Remote- und worker-lokalen M6-Pfade.

Der worker-lokale Pfad benötigt einen statt zwei RPCs und senkt den Median der serialisierten Request- und Response-Körper von 14.469 auf 8.098 Bytes. Die Reduktion von 44,032069 Prozent übertrifft das vorregistrierte 35-Prozent-Gate. Alle 30 Paare liefern bitgenau denselben Endhash; 120/120 Expertenausführungen laufen auf der Hardware-GPU und bestehen 120 unabhängige Controller-Referenzprüfungen. Das Prozess-RSS wächst nach dem Warm-up nur um 159.744 Bytes auf node1 und 49.152 Bytes auf node2.

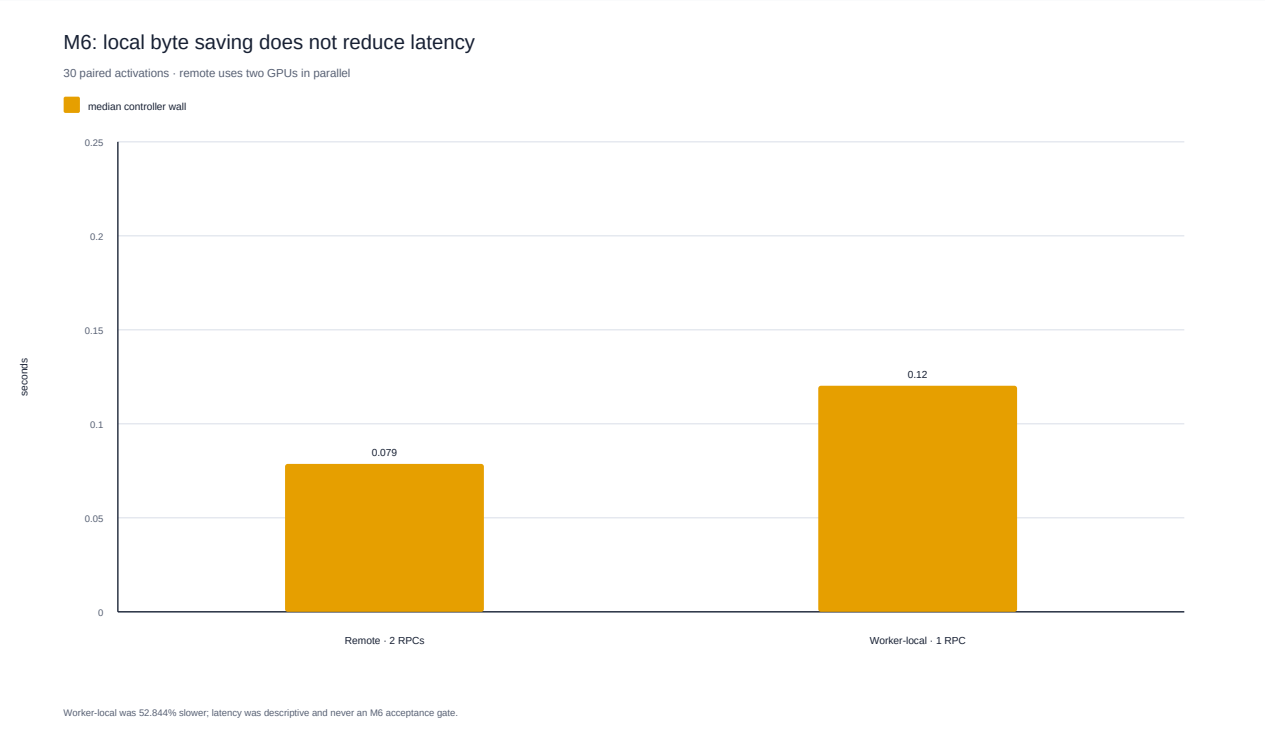

Abbildung 13. Controller-Wandzeit derselben M6-Pfade; die Byte-Ersparnis ist kein Latenzgewinn.

Die Kommunikationsreduktion wird nicht zu einer Latenzreduktion. Der lokale Median beträgt 0,120258 statt 0,078680 Sekunden und ist damit 52,843798 Prozent höher. Der Remote-Pfad nutzt zwei GPUs parallel; der lokale Pfad führt beide Experten auf einer GPU seriell aus. M6 validiert somit die Data-Plane- Voraussetzung für lokale Co-Access-Paare, nicht einen allgemeinen Geschwindigkeitsgewinn.

7.7 EXP-006: realer Routertrace und negatives Platzierungsgate

Der isolierte CPU-Lauf verarbeitet 16 eingefrorene Prompts mit 502 Tokens über 24 MoE-Schichten. Daraus entstehen 12.048 Layer-Token-Routerereignisse und 96.384 geordnete Top-8-Expertenauswahlen. Alle drei Wiederholungen besitzen dieselben Token-, Experten-, Rohlogit- und kanonischen Trace-Hashes. Das Prozess-RSS erreicht maximal 6.363.566.080 Bytes; Swap bleibt null, und die Endtemperatur von node3 beträgt 68 Grad Celsius.

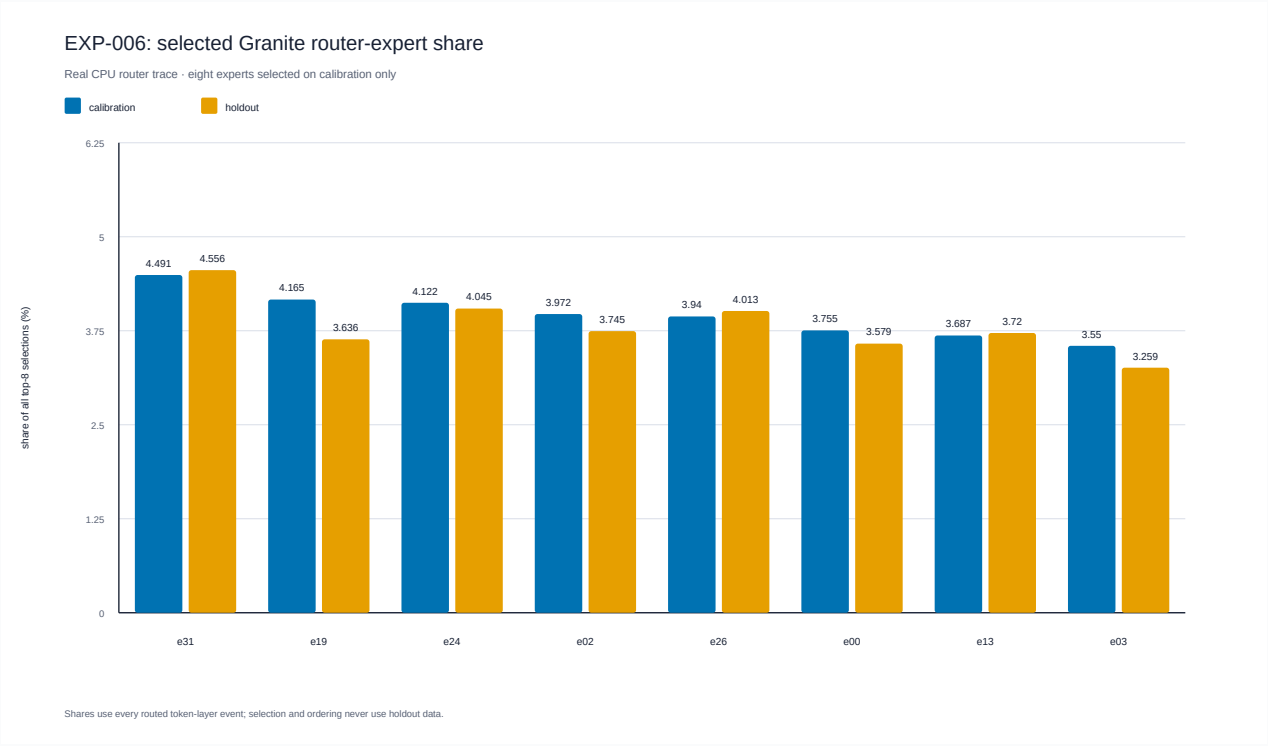

Abbildung 14. Kalibrierungsanteil der acht nach der eingefrorenen Regel ausgewählten Granite-Experten.

Die Kalibrierung wählt nach der eingefrorenen Häufigkeitsregel die Experten 31, 19, 24, 02, 26, 00, 13 und 03. Der heißeste Paarzugriff verbindet Experte 03 und 19 mit 958 Kalibrierungs- und 524 Holdout-Vorkommen.

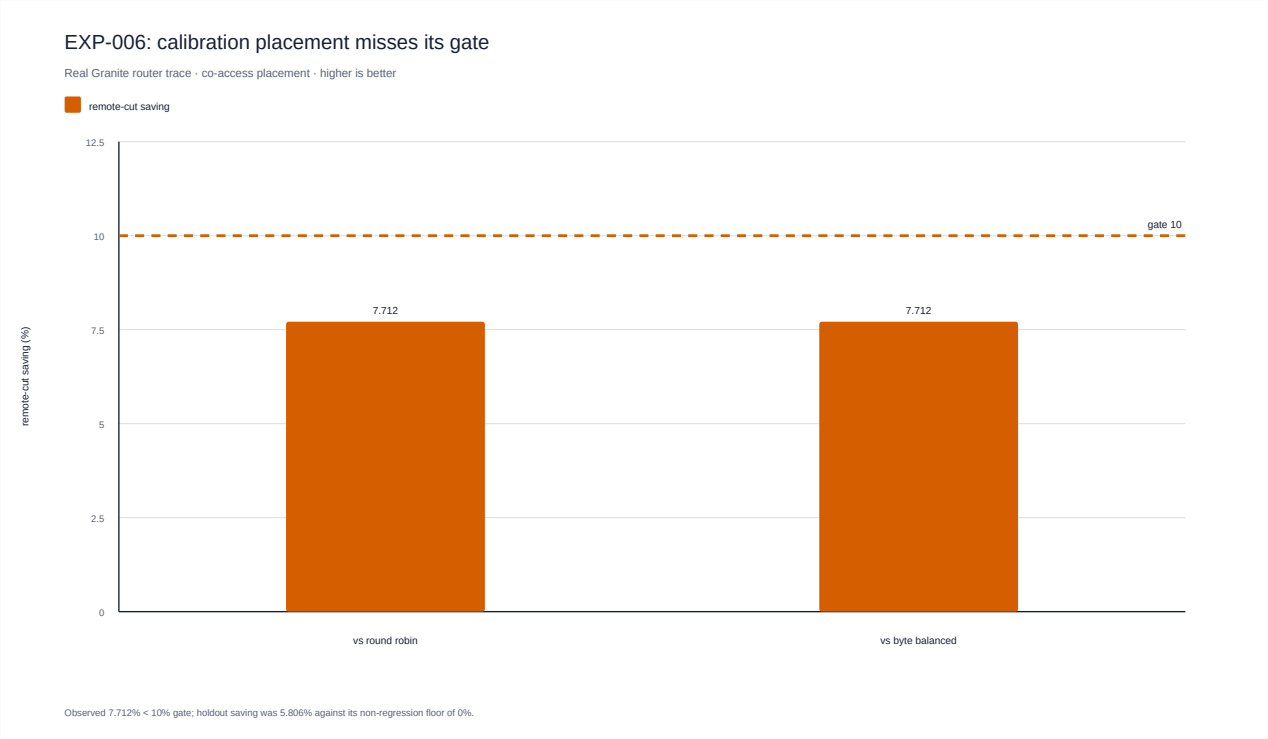

Abbildung 15. EXP-006-Kalibrierung: 7,711811 Prozent Einsparung bleiben unter dem 10-Prozent-Gate.

Round Robin und Byte Balancing erzeugen in der Kalibrierung jeweils 14.199 Remote-Paarinzidenzen; Co-Access Cost erzeugt 13.104. Die Einsparung von 7,711811 Prozent bleibt sichtbar unter der präregistrierten 10-Prozent-Schwelle. Im unangetasteten Holdout sinkt die Zahl von 9.628 auf 9.069, entsprechend 5,805983 Prozent. Auf der Uniformkontrolle enden alle Strategien bei 2.100.

Das Ergebnis ist negativ, obwohl Trace-, Kapazitäts-, Determinismus-, Holdout-, Uniform- und Fail-closed-Gates bestehen. Genau die beiden Kalibrierungsgates gegen Round Robin und Byte Balancing schlagen fehl. Deshalb starten weder Security- und Queue-Probe noch GPU-Canary oder offizieller Replay.

7.8 EXP-006B: Domänenkonditionierung erklärt den kleinen Effekt nicht

EXP-006B verarbeitet 48 neue englische Prompts aus sechs vorab bezeichneten Domänen, jeweils mit vier Kalibrierungs- und vier Holdout-Prompts. Zwei vollständige CPU-Prompt-Forward-Wiederholungen liefern bitgenau dieselben 1.681 Tokens, 40.344 Layer-Token-Routerereignisse, 322.752 geordneten Top-8-Auswahlen und vollständigen float32-Routerlogits. Peak-RSS beträgt 6.361.919.488 Bytes, Swap bleibt null und node3 endet bei 70 Grad Celsius.

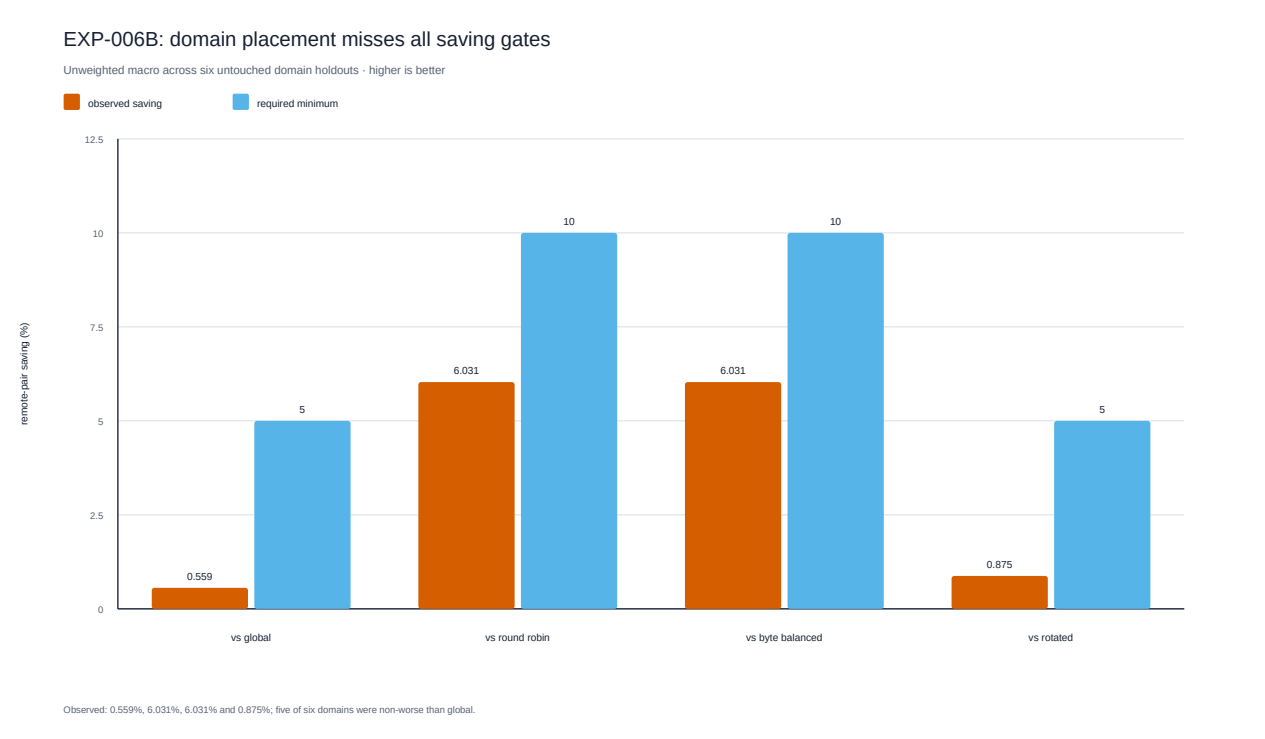

Abbildung 16. EXP-006B: Alle vier domänenspezifischen Makro-Einsparungen bleiben unter ihren vorab festgelegten Grenzen.

Die sechs ausschließlich aus ihrer jeweiligen Kalibrierung gelernten Platzierungen sparen im ungewichteten Holdout-Mittel 0,558882 Prozent gegenüber einer globalen Co-Access-Platzierung, 6,031028 Prozent gegenüber Round Robin und Byte Balancing sowie 0,875110 Prozent gegenüber der festen rotierten Domänenkontrolle. Gefordert waren 5, 10, 10 und 5 Prozent. Fünf von sechs Domänen sind nicht schlechter als global; dieses Vier-von-sechs-Gate besteht. Alle vier Spargates scheitern jedoch.

Der erste isolierte Start scheiterte noch vor Modellladen, weil unter der strikten Dateisystemgrenze kein beschreibbares temporäres Verzeichnis existierte. Dieser Fehler ist als eigener Beleg erhalten. Nur `TMPDIR` und `HF_HOME` wurden im bereits präregistrierten transienten Pfad ergänzt; Code, Korpus, Modell, Schwellen und Auswertungsreihenfolge blieben unverändert. EXP-006B startete keine GPU- oder Live-Phase.

7.9 EXP-006R: exakte Hostreplikation auf node1

EXP-006R friert den alten getesteten Commit `8ef4708`, dieselbe Runtime, dieselbe Modellrevision, denselben 16-Prompt-Korpus und alle ursprünglichen Referenzwerte ein. Ein vor dem Lauf korrigierter Speicherpfad verschiebt nur die transienten Dateien von node1s RAM-basiertem `/tmp` nach festplattenbasiertem `/var/tmp`; Modell, Code, Schwellen und Analyse bleiben unverändert.

Der einmalige isolierte Lauf liefert erneut 502 Tokens, 12.048 Routerereignisse und 96.384 Auswahlen. Token-IDs, geordnete Experten, Top-8-Scores, kanonischer Trace und vollständige float32-Routerlogits stimmen in allen drei Wiederholungen bitgenau mit node3 überein. Auch die acht ausgewählten Experten, alle Platzierungen, Remote-Paarinzidenzen und Gates sind identisch. Kalibrierung und Holdout bleiben bei 7,711811 und 5,805983 Prozent.

Die transiente Unit selbst nutzt 0 B Swap. Der Host erreicht 4.329.472 B Swapbelegung, bleibt damit deutlich unter der vorab festgelegten 64-MiB-Abbruchgrenze und endet bei 52 Grad Celsius. M3, Monitoring und Agent bleiben aktiv; FileMaker war bereits gestoppt und wurde nicht verändert.

Das Ergebnis ist eine exakte Hostreplikation, kein nachträglicher Erfolg des 10-Prozent-Gates. Es widerlegt außerdem die Vermutung, dass die Wahl des thermisch auffälligen node3 den kleinen Platzierungseffekt verursacht hat.

7.10 EXP-006C: überlappende Expertengruppen bestehen die Offline-Gates

EXP-006C friert vor Planner-Code vier Szenarien ein: die alte exklusive 3-Node-Referenz, vier Nodes ohne zusätzliche Slots, drei Nodes mit einem Zusatzreplikat und den Hauptfall mit vier Nodes, zwölf Slots und vier Zusatzreplikaten. Ein Paar ist lokal, sobald beide Experten auf mindestens einem gemeinsamen Node liegen.

Der Kalibrierungsplan repliziert Experten 02, 03, 19 und 31:

- node1: 00, 02, 31;
- node2: 03, 24, 26;
- node3: 02, 03, 19;
- node4: 13, 19, 31.

Zwei exakte Suchen über je 745.920 vollständige gültige Belegungen erzeugen denselben Plan. Er wird in Commit `8e353e6` vor dem Holdout festgeschrieben.

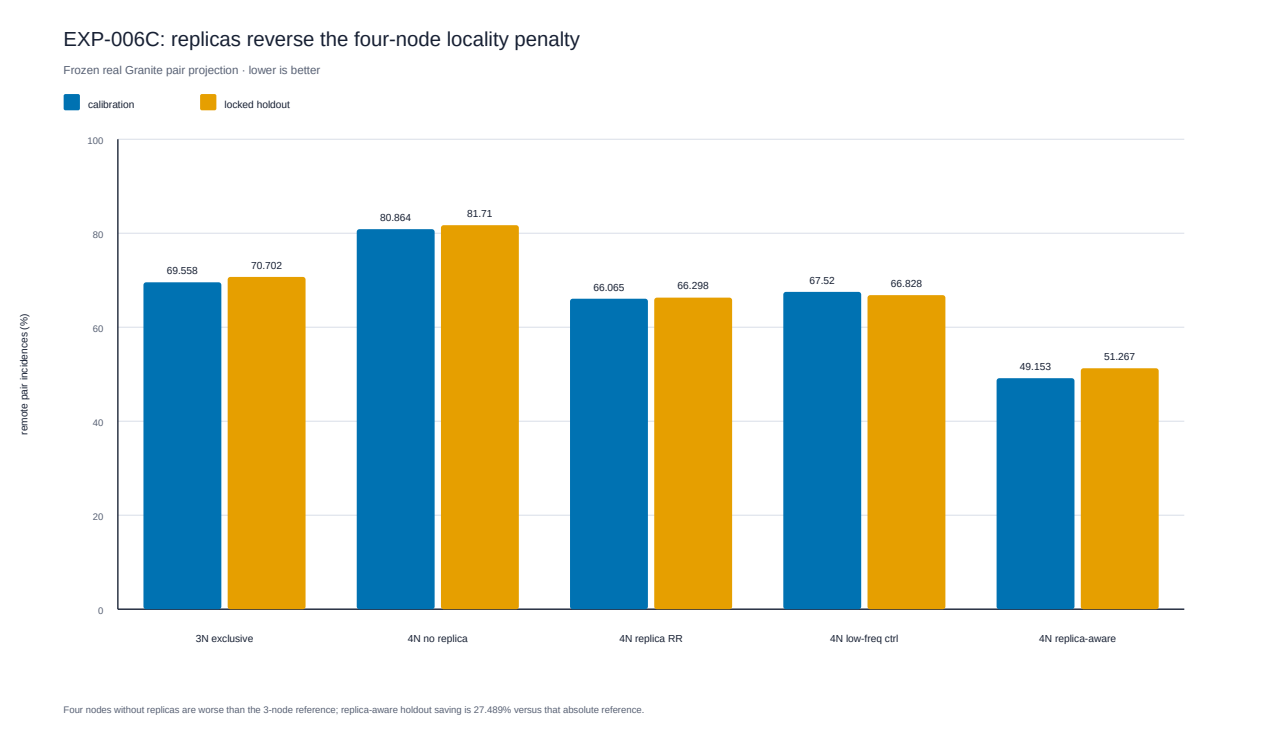

Abbildung 17. EXP-006C: Replica-aware Gruppen senken den gesperrten Holdout gegenüber der absoluten exklusiven Referenz um 27,489249 Prozent.

Der Hauptplan reduziert die Kalibrierungs-Remote-Inzidenzen gegenüber der absoluten exklusiven Referenz von 13.104 auf 9.260, also um 29,334554 Prozent. Im gesperrten Holdout sinkt die Zahl von 9.069 auf 6.576 oder um 27,489249 Prozent. Gegen kapazitätsgleiches repliziertes Round Robin und Byte-/Frequenzbalance beträgt der Holdout-Vorteil je 22,671684 Prozent; gegen die replizierte Niedrigfrequenzkontrolle 23,285114 Prozent. Alle präregistrierten Gates bestehen.

Die Kontrolle mit vier Nodes, aber nur acht nicht replizierten Slots ist schlechter als die 3-Node-Referenz: 15.234 Kalibrierungs- und 10.481 Holdout-Remote-Inzidenzen. Der Erfolg entsteht daher nicht durch eine verschlechterte Mehr-Node-Baseline.

Failover bleibt partiell. Der Verlust von node3 behält alle Experten, weil seine drei Bewohner repliziert sind. Bei Verlust von node1, node2 oder node4 fehlen Experten; betroffene Paare werden korrekt abgewiesen. Der Test belegt somit keinen allgemeinen N+1-Betrieb.

7.11 EXP-006D: der feste Plan bestätigt sich auf einem neuen Trace

EXP-006D friert vor Capture-Code und neuem Modelloutput 48 neue Prompts aus acht Domänen, Seeds 606200 bis 606247, die alten acht Experten, den EXP-006C-Hauptplan, fünf Kontrollen und alle Gates ein. Es ist weder eine neue Expertenwahl noch eine Placement-Suche zulässig.

Der einzige node1-Capture führt Attention, Router und ausgewählte Granite- Expert-MLPs zweimal vollständig auf der CPU aus. Beide Wiederholungen stimmen bei Token-IDs, geordneten Experten, vollständigen float32-Routerlogits und kanonischem Trace exakt überein. Die 1.633 Tokens erzeugen 39.192 Routerereignisse und 89.506 projizierte Paarinzidenzen über alle 28 möglichen Paare der festen Expertengruppe.

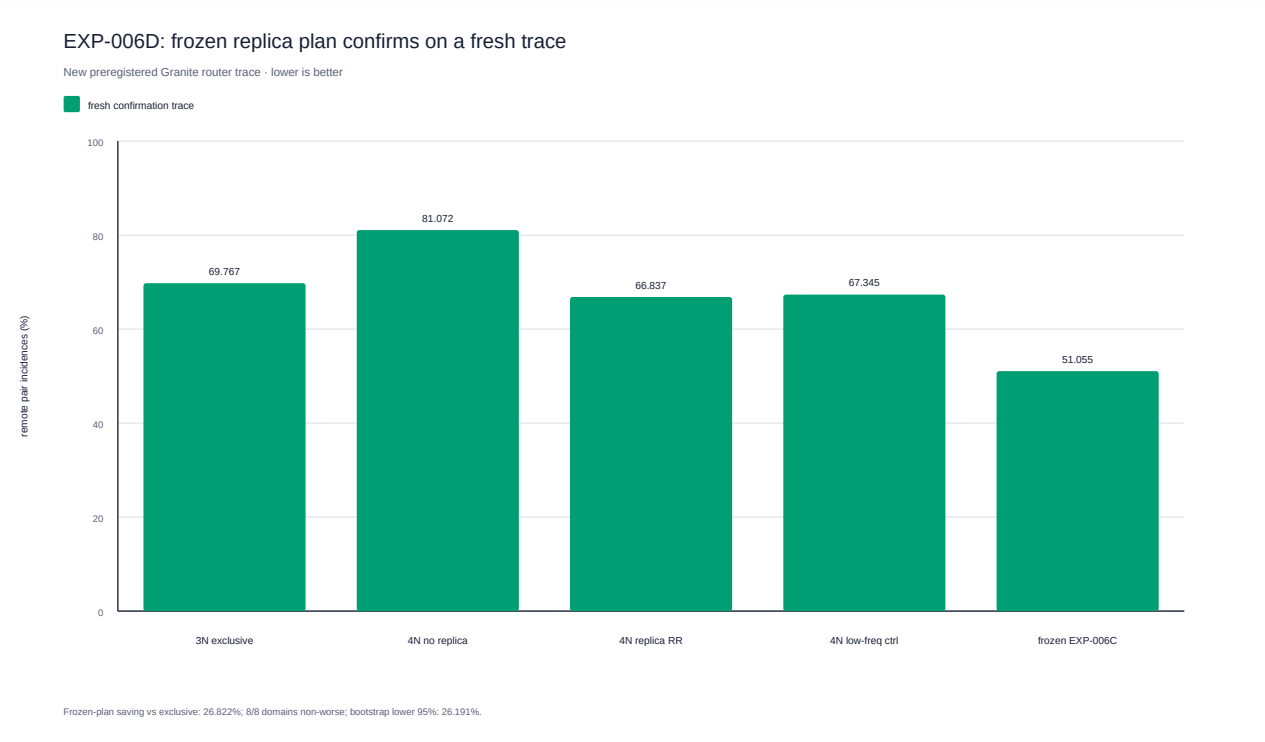

Abbildung 18. EXP-006D: Der unveränderte Replica-aware Plan senkt Remote-Paarinzidenzen auf einem neuen Routertrace gegenüber der exklusiven Referenz um 26,821574 Prozent.

Auf der neuen Spur sinken die Remote-Paarinzidenzen von 62.446 bei der exklusiven 3-Node-Referenz auf 45.697 beim unveränderten Plan. Das entspricht 26,821574 Prozent. Repliziertes Round Robin und Byte-/Frequenzbalance enden jeweils bei 59.823; die Niedrigfrequenzkontrolle bei 60.278. Der Kandidat ist damit 23,612992, 23,612992 und 24,189588 Prozent besser.

Das ungewichtete Acht-Domänen-Mittel beträgt 26,832559 Prozent, und alle acht Domänen sind gegenüber der exklusiven Referenz nicht schlechter. Die untere 95-Prozent-Grenze des 10.000-fachen Prompt-Bootstraps liegt bei 26,191329 Prozent. Alle vorab festgelegten Bestätigungsgates bestehen.

Die 4-Node/8-Slot-Kontrolle ohne Replikate bleibt mit 72.564 Remote- Inzidenzen schlechter als die 3-Node-Referenz. Der Erfolg ist daher erneut kein Effekt der Nodezahl allein. Die neue Stichprobe stärkt den Offline-Planungsbefund, führt aber weder Granite auf den Intel-GPUs aus noch misst sie Netzwerkbytes, Latenz, Energie oder Produktions-Failover.

7.12 EXP-006E: Schutzgate stoppt vor der Live-Metrik

EXP-006E bildet den logischen vierten Slot auf den physischen node8 ab und bindet Referenz, Kandidat, 28 Paarhäufigkeiten, 112 gepaarte Fälle, 448 vorgesehene GPU-Expertenausführungen und das 7,5-Prozent-Bytegate vor dem Live-Lauf. Die Implementierung besteht auf Node1 195/195 Vertragstests.

Node8 erfüllt beim letzten Preflight Hostname, RAM-, Swap-, Load-, Temperatur-, Rendergeräte-, Voice-Health- und Dienstgates. Der gleichzeitig geschützte CT333 meldet jedoch:

```
pressurecpufull=0.28
pressureiofull=0.00
pressurememoryfull=0.00
swap=0
```

Die Präregistrierung verlangt für alle drei Full-Pressure-Werte exakt 0,00. Damit ist der einzig zulässige Ausgang `node8_ineligible_before_live_execution`. Die übrige EXP-006E-Flotte darf nicht starten.

Größe	Beobachtung
offizielle Node8-Qualifikationen	0
offizielle Vier-Domain-Replays	0
GPU-Expertenausführungen	0
Strategie-Paarergebnisse	0
Anwendungsbyte-Ersparnis	nicht gemessen

Zwei frühere Prozessstarts erreichten wegen `PrivateTmp` keinen Worker; zwei Monitorstarts endeten an einer zu kurzen Diagnosezeit beziehungsweise einer nicht präregistrierten Speichergrenze. Alle vier Diagnosen liegen vor jeder offiziellen Qualifikation und erzeugen null GPU-Jobs. Sie werden berichtet, aber nicht als Messwiederholungen umgedeutet.

Im Postflight laufen CT333, CT808, Voice-Dienste und die produktiven M3-Dienste unverändert; CT333 meldet wieder 0,00 Full-Pressure und null Swap. Dieses spätere Grün macht den eingefrorenen Startbefund nicht rückwirkend gültig. EXP-006E belegt somit die priorisierte Schutzlogik. Es beantwortet weder die Byte- noch die Latenz- oder GPU-Transferfrage.

8. Diskussion

8.1 Was EXP-005 für ein Model OS bedeutet

Das Ergebnis rechtfertigt eine klare Scheduler-Schnittstelle: Modellgraph, Expertengrößen, Co-Access-Häufigkeiten und Node-Grenzen müssen gemeinsam in die Platzierung eingehen. Gleich viele Experten oder gleiche Bytes allein sind unzureichend.

Die Kapazitätskurve zeigt außerdem, dass kleine Aufrüstungen sprunghaft wirken können. Zusätzlicher RAM oder eine bessere Repack-Grenze ist nicht nur „etwas mehr Kapazität“, sondern kann eine vollständige Affinitätsgruppe lokal halten und dadurch einen Kommunikationssprung vermeiden.

8.2 Von der simulierten Co-Location zur Live-Data-Plane

M6 schließt die technische Lücke von EXP-005 für einen kleinen, synthetischen Zwei-Experten-Job: Eine gemeinsame Aktivierung wird einmal übertragen, beide Experten rechnen weiter auf der GPU, und der Worker gibt nur das zusammengeführte Ergebnis plus kompakte Prüfquittungen zurück.

Der Befund trennt zwei Schedulerziele, die nicht gleichgesetzt werden dürfen. Lokale Co-Access-Platzierung spart Kommunikationsvolumen und kann knappe Netzkapazität freihalten. Parallel verteilte Ausführung kann bei genügend GPUs trotzdem schneller sein. Ein künftiges Model OS muss deshalb Bytekosten, GPU-Parallelität, Queuezustand und Latenzziel gemeinsam bewerten.

8.3 Rolle der CPUs

Die GPUs führen die zugelassenen Expert- und GEMM-Arithmetik aus. CPUs bleiben parallel für:

- Netzwerk und TLS;
- Pufferverwaltung;
- De-/Quantisierung;
- RAM-Befüllung und Freigabe;

- Referenzprüfung und Telemetrie.

M4 zeigt, dass begrenzte CPU-/RAM-Arbeit neben korrekten GPU-Routen möglich ist, allerdings mit 13,158 Prozent höherer Median-Wandzeit. Diese Kopplung muss in spätere Kostenmodelle eingehen.

8.4 Transfer vom synthetischen zum realen Trace

EXP-006 relativiert den großen EXP-005-Effekt. Auf dem konstruierten EXP-005-Trace konnten ganze Affinitätsgruppen lokal gehalten werden; beim realen Granite-Router verteilt sich die ausgewählte Top-8-Co-Access-Struktur breiter. Die exhaustive Suche findet weiterhin einen reproduzierbaren Vorteil, aber keinen ausreichend großen.

Der positive Holdout-Wert und die neutrale Uniformkontrolle sprechen gegen einen offensichtlichen Optimierungsfehler. Sie rechtfertigen jedoch keine nachträgliche Absenkung des Gates. Für künftige Scheduler folgt daraus, dass Co-Access-Konzentration zunächst gemessen werden muss und Placement nur ein Ziel neben Parallelität, Queuezustand, Aktivierungsgröße und Latenz sein kann.

8.5 Domänenlabels sind hier kein ausreichendes Schedulersignal

EXP-006B prüft eine konkrete Erklärung für den kleinen EXP-006-Effekt und verwirft sie unter den eingefrorenen Bedingungen. Domänenspezifische Kalibrierung erzeugt zwar in fünf von sechs Domänen keine Verschlechterung, aber nur einen sehr kleinen mittleren Vorteil gegenüber einer globalen Platzierung. Auch die falsch rotierte Kontrolle bleibt zu nahe am Kandidaten.

Das spricht nicht gegen jede Form von Workload-Konditionierung. Es zeigt enger, dass grobe redaktionelle Domänenlabels bei vier Kalibrierungsprompts je Domäne keine ausreichend stabile zusätzliche Co-Access-Struktur liefern. Ein späterer Scheduler sollte deshalb Routerstatistik, Aktivierungsgröße, Parallelitätswert und Umschaltkosten gemeinsam untersuchen, statt Themenlabels als starken Proxy vorauszusetzen.

8.6 Replizierte Expertengruppen als Model-OS-Funktion

EXP-006C bestätigt den vorher fehlenden Planungsmechanismus innerhalb seiner Offline-Grenze. Häufig gemeinsam geroutete Experten können auf mehreren Nodes resident sein, sodass ein Experte gleichzeitig zwei unterschiedliche lokale Gruppen verbindet. Der exakte Scheduler entscheidet gemeinsam über Replikatwahl, Gruppenabdeckung, Slots, modellierte Last, node3-Thermik und Fail-closed-Ausfälle.

Die 4-Node/8-Slot-Kontrolle ist für die Interpretation zentral. Eine reine 2/2/2/2-Verteilung schneidet mehr Paare als der alte 3-Node-Plan. Zusätzliche Nodes helfen erst, wenn sie zusätzliche, gezielt platzierte Kopien aufnehmen. Model OS muss daher Nodezahl, Replikatbudget und Paarabdeckung gemeinsam optimieren; „mehr Nodes“ ist kein eigenständiges Schedulerziel.

Der positive EXP-006C-Holdout war code-seitig sauber vom Plan getrennt, aber nicht investigator-blind neu: Seine alten Gesamtergebnisse waren aus EXP-006 bekannt. EXP-006D schließt diese konkrete Bestätigungslücke mit einem vor dem Modelllauf eingefrorenen Korpus und dem unveränderten Plan. Die breite Einsparung über acht Domänen und die positive Bootstrap-Untergrenze sprechen gegen einen günstigen Einzelholdout.

Die Bestätigung bleibt dennoch lokal erstellt und offline. Sie qualifiziert weder node4 als Live-Worker noch Granite-Expertengewichte auf den Intel-GPUs. Der nächste sinnvolle Schritt ist deshalb keine weitere nachträgliche Planoptimierung, sondern eine separat präregistrierte, begrenzte GPU-/Data-Plane-Prüfung mit expliziten Sicherheits-, Ressourcen- und Abbruchgrenzen.

8.7 Ein Schutzgate ist ein Ergebnis, aber kein Leistungsbefund

EXP-006E zeigt eine für Model OS zentrale Priorität: Der Scheduler darf geschützte Dienste nicht für eine nachgelagerte Forschungsmetrik belasten. Das gilt auch dann, wenn alle statischen Hardwaremerkmale günstig aussehen und das Pressure-Signal im Postflight wieder verschwindet. Die vorher festgelegte Entscheidungsregel muss am beobachteten Startpunkt gelten.

Der Ausgang darf zugleich nicht überinterpretiert werden. Er zeigt weder, dass node8 dauerhaft ungeeignet ist, noch dass der replizierte Plan keine Anwendungsbytes spart. Ein einzelner Full-Pressure-Wert ist kein Leistungsmodell. Er ist hier ein konservativer Betriebsindikator mit Vorrang.

Ein Nachfolger darf die Nullgrenze nicht nach Sichtung dieses Ergebnisses einfach lockern und sich weiterhin EXP-006E nennen. Wissenschaftlich sauber wären etwa ein eigener vierter Failure-Domain-Host ohne geschützte Last oder eine neue Präregistrierung mit einer vorab begründeten, zeitlich wiederholten Eligibility-Regel. Erst danach ist die offene Bytefrage wieder messbar.

9. Bedrohungen der Validität

Interne Validität

- EXP-005 verwendet absichtlich konstruierte synthetische Traces.
- Der Byte-Anker stammt aus einem kleinen JSON/mTLS-Canary, nicht aus einem produktiven MoE-Stack.
- M6 vergleicht zwei parallele Remote-GPUs mit zwei seriellen lokalen GPU-Ausführungen; sein Latenzunterschied isoliert nicht den Transport allein.
- EXP-006 verwendet einen festen kleinen Prompt-Korpus. Die beobachtete Routerverteilung kann von Domäne, Sprache, Sequenzlänge und Tokenisierung abhängen.
- EXP-006B verwendet ebenfalls nur vier Kalibrierungs- und vier Holdout-Prompts je Domäne. Die Domänen sind redaktionelle Kategorien und keine beweisbare latente Routerklassifikation.
- EXP-006 wählt die acht Experten nach Kalibrierungshäufigkeit; andere vorab definierte Auswahlziele könnten andere Platzierungen ergeben.
- EXP-005, EXP-006 und EXP-006R erlauben keine Replikate oder überlappenden Expertengruppen; ihre Optimalität gilt nur für exklusive Platzierung.
- EXP-006C verwendet den alten EXP-006-Holdout. Er ist gegenüber dem neuen Planner gesperrt, aber seine früheren Summen waren den Forschenden bekannt; er ersetzt keinen neuen investigator-blinden Trace.
- EXP-006D verwendet einen neuen, vor dem Modelllauf hashgesperrten Korpus, seine Prompttexte wurden aber lokal verfasst und nicht durch eine externe Partei verborgen oder repliziert.
- EXP-006D bestätigt nur die acht bereits in EXP-006 ausgewählten Experten; es prüft keine neue Expertensuche und keine Übertragbarkeit auf andere MoE-Modelle.
- EXP-006E beobachtet am entscheidenden Preflight-Punkt genau einen negativen CT333-CPU-Full-Pressure-Wert. Das genügt für das präregistrierte Schutzgate, beschreibt aber keine zeitliche Pressure-Verteilung.
- EXP-006E erzeugt keine offizielle Qualifikation und keine Data-Plane-Beobachtung. Aus seinem Ausgang kann weder ein positiver noch ein negativer Byte-, GPU- oder Latenzeffekt geschätzt werden.
- Die präqualifikatorischen Startdiagnosen zeigen reale Deployment-Reibung, liegen aber vor jedem Worker- und GPU-Output und sind keine Leistungsstichprobe.
- EXP-006Cs residente Bytes stammen aus den begrenzten synthetischen M6-Bündeln und sind keine realen Granite-Expertengewichte.
- Kleine Zeitunterschiede hängen von Queue- und Betriebssystemzustand ab.
- Mehrere Hardwaremessungen sind Mediane weniger Wiederholungen.

Externe Validität

- Die GPUs sind Intel HD Graphics 4000 und nicht repräsentativ für moderne Beschleuniger.
- Das physische Netzwerk ist 1 Gb/s.
- Der Expertentensor ist 16×16 und der Graph hat drei Schichten.
- Beim realen Open-Weight-MoE wurde ein vollständiger CPU-Prompt-Forward einschließlich ausgewählter Expert-MLPs ausgeführt. Nicht ausgeführt wurden autoregressive Generation, verteilte End-to-End-Inferenz oder Granite-Rechenpfade auf den Intel-GPUs.
- Node8s Intel Iris Graphics 5100 unterscheidet sich von den drei HD-Graphics-4000-Workern. EXP-006E erzeugt jedoch keine Messung, mit der dieser Unterschied bewertet werden könnte.

Konstruktvalidität

- „Remote-Schnitt“ ist ein Graphmaß, keine vollständig beobachtete Netzwerksession; M6 misst JSON-Anwendungsbytes, nicht Wire-Bytes. Die EXP-006-Bytetotale sind ausschließlich daraus modelliert.
- `pressurecpufull` ist in EXP-006E ein Schutz- und Zulassungskonstrukt, keine direkte Messung von GPU-Leistung, Netzwerklast oder Modellqualität.

- Geschätzte GFLOP/s beziehen sich auf den festen Shader.
- Zwei-Stream-RAM-Skalierung ist kein direkter Speicherkanalzähler.
- SSD-Kapazität darf nicht mit interaktiver Modellresidenz verwechselt werden.

10. Reproduzierbarkeit

Alle relevanten Artefakte sind versioniert:

- Präregistrierung EXP-005: Commit `5d88834`;
- Simulator und Vertragstests: Commit `7e5c1c5`;
- akzeptierte M5.1-Basis: Commit `b50c2a1`;
- Präregistrierung M6: Commit `ede8b29`;
- M6-Implementierung vor dem Live-Lauf: Commit `5642e04`;
- Präregistrierung EXP-006: Commit `8ee5a5c`;
- EXP-006-Implementierung: Commit `8c12b2f`;
- enge CPU-Distributionskorrektur: Commit `8ef4708`;
- Präregistrierung EXP-006B: Commit `b1b71f8`;
- Korrektur der CPU-Ausführungsgrenze vor dem EXP-006B-Lauf: Commit `3b94d7a`;
- EXP-006B-Implementierung vor dem Modelllauf: Commit `9db35b9`;
- EXP-006R-Präregistrierung: Commit `d9f9960`;
- EXP-006R-Speicherpfadkorrektur vor Modellladen: Commit `45eafb5`;
- EXP-006R-Ergebniscommit: `20c4a55`;
- EXP-006C-Präregistrierung vor Code und Auswertung: Commit `c9e724b`;
- EXP-006C-Implementierung: Commit `867962d`;
- EXP-006C-Kalibrierungs-Freeze vor Holdout: Commit `8e353e6`;
- EXP-006C-Ergebniscommit: `0e86d71`;
- EXP-006D-Präregistrierung vor Code und Modelloutput: Commit `8315c78`;
- EXP-006D-Implementierung vor dem Modelllauf: Commit `e54f638`;
- EXP-006D-Ergebniscommit: `dc093b6`;
- EXP-006E-Präregistrierung vor Implementierung und Live-Ausführung: Commit `fd18fe1`;
- EXP-006E-Implementierung vor jedem Live-Output: Commit `964a0fb`;
- enger read-only Timeout-Fix vor jedem Live-Output: Commit `5601d52`;
- EXP-006E-Ergebniscommit: `19392f7`;
- abschließende EXP-006E-Provenienzdokumentation: Commit `4fdc85a`;
- maschinenlesbare Evidenz unter `model-os/evidence/`;
- deterministische SVG-Master unter `model-os/docs/publication/figures/`;
- vollständige Quellen- und Ausgabehashes in `figure-manifest.json`.

Die aktuelle geheimnisfreie Linux-Suite umfasst nach EXP-006E 195 Tests. Sie prüft neben Runtime-, Sicherheits-, Placement- und Pufferverträgen auch, dass jede Evidenzdatei eine Publikationsentscheidung besitzt und zwei unabhängige Abbildungs-Builds aller 21 SVG-Master bitidentische Hashes erzeugen. Die EXP-006-Tests prüfen zusätzlich Modell-/Runtime-Identität, Traceintegrität, Kalibrierungsisolation und das downstream Fail-closed-Verhalten. EXP-006B ergänzt Korpus-, Domänen-, Platzierungsfreeze-, Rotationskontroll- und Holdout-Isolationstests. EXP-006R ergänzt die drei vorab erklärten Cross-Host-Ergebnisarten und trennt fachliche Replikation von vollständiger Rohlogit-Identität. EXP-006C ergänzt Replikat-Lokalität, exakte Vier-Node-Suche, kapazitätsgleiche Kontrollen und Ausfallklassifikation. EXP-006D ergänzt einen neuen Promptkorpus, exakte Doppel-Captures, feste Planauswertung, acht Domänen und einen deterministischen Prompt-Bootstrap. EXP-006E ergänzt Quell- und Placement-Bindung, acht verschiedene Expertensignaturen, Drei-Resident-Worker, Voice-/Pressure-Schutz, Queue- und Sicherheitsgrenzen, einen festen Paarreplay und die fail-closed

Ergebnisintegration. Das exakte geheimnisfreie `867962d`-Archiv mit SHA-256 `aea5da0425d92cef2f27935d0829eec9491c228e5a87fc35d2e13ba0ba90b400` bestand 165/165 Tests auf node1 Linux in 133,653 Sekunden. Das exakte positive Ergebniscommit-Archiv `0e86d71` mit SHA-256 `96100d0dcdd0a50c347503a9ab6cc6986999191b61ad1fc388147b2fcbecc2f5` bestand dieselben 165/165 Tests in 132,233 Sekunden. Das exakte positive EXP-006D-Ergebniscommit-Archiv `dc093b6` mit SHA-256 `45d709c710056f3e92eb87fe0c93bcd1e8f34db9546340554cbbe7112187c07f` bestand auf node1 176/176 Tests in 163,841 Sekunden. Das exakte EXP-006E-Ergebniscommit-Archiv `19392f7` mit SHA-256 `a659065c0fe652480ea18215485d8ea0b4b009d2541d5bb6baa226635f8f286d` bestand auf node1 195/195 Tests in 166,896 Sekunden. Sein maschinenlesbares Ergebnismanifest hat SHA-256 `ff427f8064ea90c7b49257df9d0189a31333e5b15703ddd9b92ad4e9db6c7c8f`.

11. Nächste Experimente

Vor einer belastbaren Systempaper-Einreichung sind mindestens notwendig:

1. die durch EXP-006E offen gebliebene Vier-Failure-Domain-Bytefrage nur in einem separat präregistrierten Nachfolger untersuchen: bevorzugt auf einem dedizierten vierten Host ohne geschützte Last oder mit einer vorab begründeten wiederholten Eligibility-Regel;
2. reale Expert-MLPs und versteckte Aktivierungen auf den Intel-GPUs nur unter neuen, vorab bestandenen Ressourcen- und Platzierungsgates ausführen;
3. JSON-Anwendungsbytes mit einer kontrollierten Ethernet-Wire-Byte-Messung und Energieaufnahme ergänzen;
4. Linkmatrix und Wärme über längere, kontrollierte Läufe erfassen;
5. vollständige Replikation aller Experten und Node-Ausfall als separates EXP-007 präregistrieren;
6. verwandte MoE-Systemarbeiten systematisch ergänzen;
7. mindestens eine externe Replikation oder unabhängige Artefaktprüfung durchführen.

12. Schlussfolgerung

Model OS zeigt auf einem absichtlich knappen Commodity-Cluster zwei zusammengehörige, aber getrennte Fortschritte plus eine live gemessene Brücke. Erstens kann eine stark begrenzte, auditierbare GPU-Data-Plane residente Experten und einen kleinen mehrschichtigen Graphen korrekt und parallel ausführen. Zweitens kann eine modellbewusste Offline-Platzierung auf eingefrorenen synthetischen Traces deutlich weniger Co-Access-Schnitte finden als Round Robin und reine Byte-Balancierung. M6 zeigt schließlich, dass worker-lokale Ausführung die Anwendungsbytes bei bitgleichem Ergebnis real reduziert, dabei aber langsamer als der parallele Zwei-GPU-Pfad bleibt.

EXP-006 ergänzt eine notwendige Gegenprobe: Ein realer, deterministisch erfasster Granite-Routertrace bestätigt einen kleinen Co-Access-Vorteil, aber nicht die präregistrierte Mindeststärke. Der daraufhin unterlassene GPU-Replay ist Teil des Ergebnisses. Damit trennt das Projekt eine belastbare Routermessung von einer nicht belegten Behauptung über verteilte MoE-Inferenz.

EXP-006B grenzt die Erklärung weiter ein: Grobe, vorab vergebene Promptdomänen erzeugen auf einem neuen Korpus keinen hinreichend großen zusätzlichen Placement-Vorteil. Der Befund ist erneut negativ, aber entscheidungsrelevant: Weder das 10-Prozent-Gate des ursprünglichen Experiments noch die neue Fünf-Prozent-Grenze gegenüber globaler Kalibrierung darf nachträglich abgesenkt werden.

EXP-006R zeigt zusätzlich, dass das negative EXP-006-Ergebnis auf einem zweiten Host bis zu den Rohlogits bitgenau reproduzierbar ist. Der Befund betrifft jedoch nur die bisher implementierte exklusive Platzierung. Replikate und überlappende Expertengruppen sind keine nachträgliche Widerlegung, sondern ein separat präregistrierter Model-OS-Planungsraum.

EXP-006C zeigt in diesem Planungsraum einen deutlich größeren Offline-Effekt. Entscheidend ist nicht die vierte Nodezahl allein - ohne zusätzliche Slots wird sie schlechter -, sondern die gezielte Mehrfachnutzung häufig verbindender Experten. EXP-006D bestätigt den unveränderten Plan anschließend auf einem neuen, vorab eingefrorenen Routertrace: 26,821574 Prozent gegenüber der exklusiven Referenz, acht von acht bessere Domänen und eine klar positive Bootstrap-Untergrenze. Damit ist die vorher benannte Offline- Bestätigungslücke geschlossen, nicht aber die GPU- oder Live-Grenze.

EXP-006E versucht diese Live-Brücke unter einem vorab festen Vier-Failure-Domain-Vertrag. Der Versuch endet korrekt vor der ersten offiziellen Qualifikation, weil der geschützte CT333 das exakte CPU-Full-Pressure-Nullgate verletzt. Das ist ein positiver Befund über die Schutzlogik und ein unentscheidbarer Befund über die geplante Data-Plane-Wirkung. Null GPU-Jobs und eine ungemessene Bytefrage sind hier wesentliche Resultate, keine fehlenden Tabellenwerte.

Der wichtigste Befund ist nicht, dass alte Hardware plötzlich ein großes LLM trägt. Er ist, dass Speicherkapazität, Co-Access, Kommunikationskosten, Pufferbesitz, Sicherheit und Wärme als gemeinsamer Planungsraum behandelt werden können - und dass sowohl Erfolg als auch Fehlschlag reproduzierbar begrenzt werden.

Literatur- und Artefakthinweis

Die Arbeitsbibliographie liegt in `Project-Ant-Colony/03_literature/bibliography.bib`. Vor einer Einreichung wird sie um MoE-Placement- und Expert-Parallelism-Literatur erweitert und in das Zitationsformat des Zielvenues überführt.

Die vollständige Abbildungssammlung mit Quellenklassifikation befindet sich in `model-os/docs/PUBLICATION_VISUALIZATION_POLICY.md`.

Redaktionelle Restpunkte vor einer Einreichung

- Zielvenue und Format festlegen;
- optionale formale Institution und ORCID ergänzen;
- MoE-Related-Work vervollständigen;
- Tabellen für Hardware, Sicherheitsgates und Commit-Historie verdichten;
- EXP-006E-Nachfolger für die offene Vier-Domain-Bytefrage separat präregistrieren;
- Energie-/Leistungsaufnahme nachmessen;
- englische Fassung und formale Zitationen erzeugen;
- externe technische Durchsicht dokumentieren.

Anhang A: Ergänzende Messgrafiken

Die folgenden Abbildungen gehören zum aktuellen, reproduzierbaren Abbildungsbestand, werden im Haupttext aber nicht direkt benötigt.

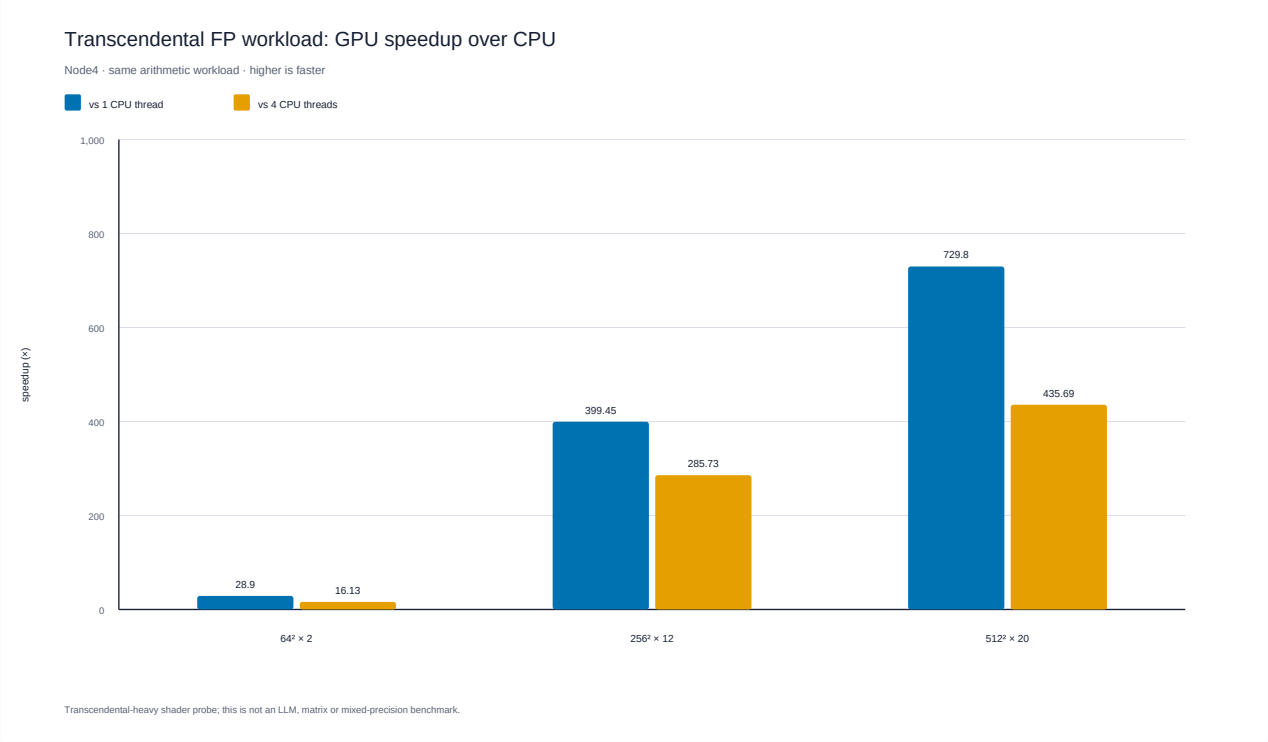

Abbildung 19. GPU-Beschleunigung einer transzendenten FP-Last gegenüber der CPU.

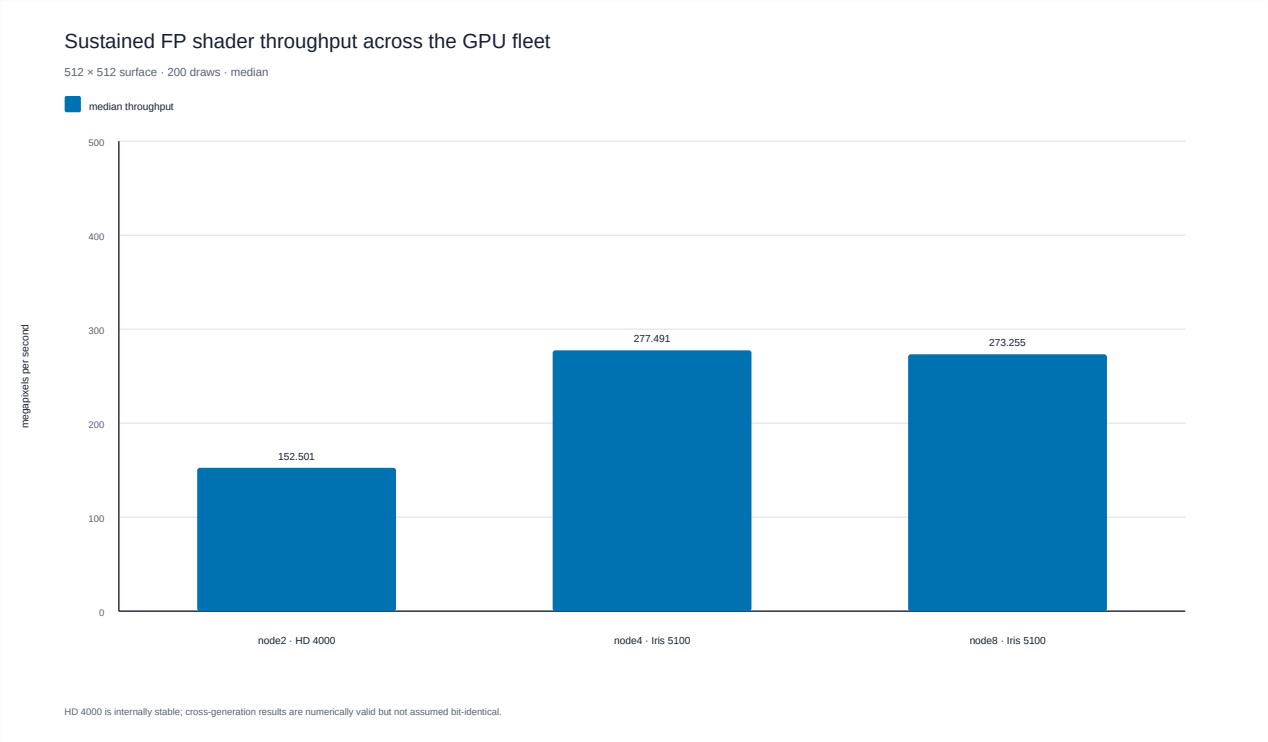

Abbildung 20. Anhaltender FP-Shader-Durchsatz im getesteten GPU-Bestand.

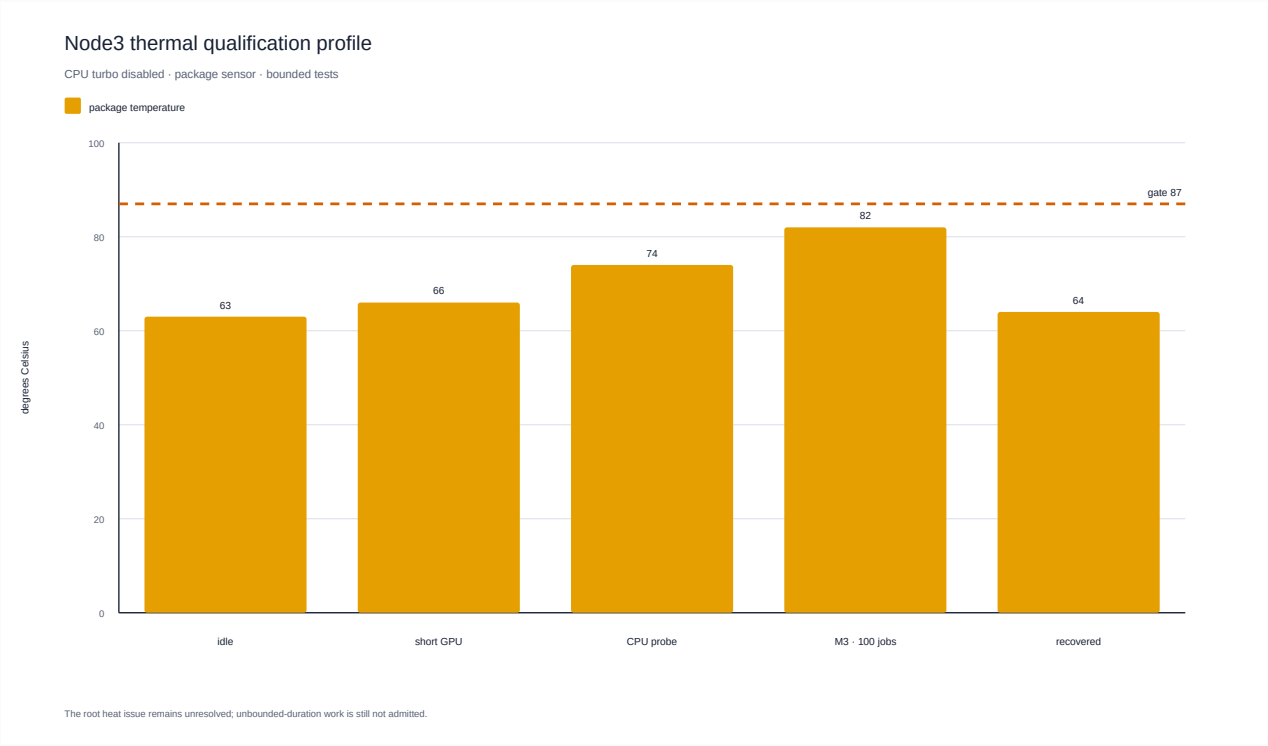

Abbildung 21. Thermisches Profil von node3 während der Qualifikationslast.

Literatur

Im PDF aufgelöste Arbeitsbibliographie der im Abschnitt „Verwandte Arbeiten“ verwendeten Quellen. Der MoE-spezifische Literaturteil bleibt vor einer Einreichung zu vervollständigen.

[1] **Chen, T. et al. (2018).** TVM: An Automated End-to-End Optimizing Compiler for Deep Learning. *13th USENIX Symposium on Operating Systems Design and Implementation, 578-594.*
https://www.usenix.org/conference/osdi18/presentation/chen

[2] **Zheng, L. et al. (2020).** Ansor: Generating High-Performance Tensor Programs for Deep Learning. *14th USENIX Symposium on Operating Systems Design and Implementation, 863-879.*
https://www.usenix.org/conference/osdi20/presentation/zheng

[3] **Xiao, W. et al. (2018).** Gandiva: Introspective Cluster Scheduling for Deep Learning. *13th USENIX Symposium on Operating Systems Design and Implementation, 595-610.*
https://www.usenix.org/conference/osdi18/presentation/xiao

[4] **Mao, H. et al. (2019).** Learning Scheduling Algorithms for Data Processing Clusters. *Proceedings of ACM SIGCOMM.*
https://web.mit.edu/decima/index.html

[5] **Stich, S. U. (2019).** Local SGD Converges Fast and Communicates Little. *International Conference on Learning Representations.*
https://openreview.net/forum?id=S1g2JnRcFX

[6] **Lin, Y. et al. (2018).** Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training. *International Conference on Learning Representations.*
https://openreview.net/forum?id=SkhQHMW0W

[7] **Fawzi, A. et al. (2022).** Discovering Faster Matrix Multiplication Algorithms with Reinforcement Learning. *Nature 610, 47-53.*
https://doi.org/10.1038/s41586-022-05172-4

[8] **Mankowitz, D. J. et al. (2023).** Faster Sorting Algorithms Discovered Using Deep Reinforcement Learning. *Nature 618, 257-263.*
https://doi.org/10.1038/s41586-023-06004-9

[9] **Real, E. et al. (2020).** AutoML-Zero: Evolving Machine Learning Algorithms From Scratch. *Proceedings of the 37th International Conference on Machine Learning, 8007-8019.*
https://proceedings.mlr.press/v119/real20a.html

Bauprovenienz

Feld	Wert
PDF-Version	Paper I · v1.0 · 26. Juli 2026
Quellmanuskript	docs/publication/MODEL_OS_RESEARCH_PAPER_DRAFT.md
Quell-SHA-256	ddb6ff31fbaacadc7006fef8c03e8cb25fb5ff27626721ee488edd1fd90dccd9
Git-Commit beim Bau	7048fa5
Abbildungen	21 SVG-Master
Abbildungsmanifest	docs/publication/figure-manifest.json

Diese PDF ist ein abgeleitetes Artefakt. Für inhaltliche Änderungen wird das Markdown-Manuskript bearbeitet und die PDF neu gebaut.