AI Ramp Accelerate: Engineering Playbook

Optimization Strategies for Ungated 70B and 700B-Class LLMs

Target: ≥40% Throughput Uplift (Tokens/s)

TENSOR™ NETWORKS

AI Ramp

Product Version: AI Ramp™ Accelerate (v196)

System Engineering White Paper / Implementation Guide

The Operational Imperative

- Data Center Constraints:
  - Power Envelopes (MW)
  - Cooling Capacity
  - GPU Availability

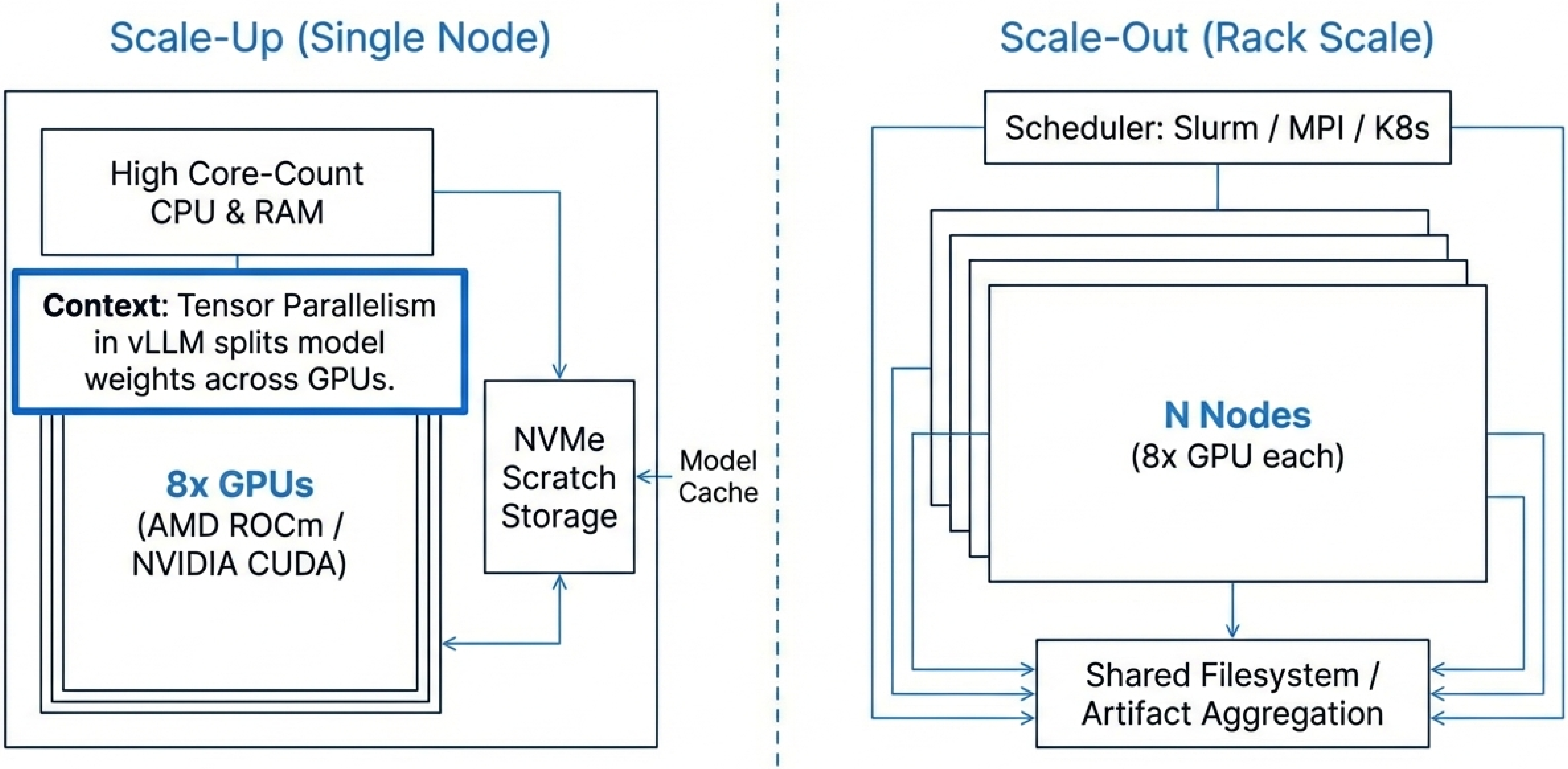
The Mission

Objective: Increase useful work per unit time (Tokens/s) and useful work per unit energy (Tokens/J) **without hardware replacement.**

Success Metrics

Primary KPI	Throughput Uplift **≥40%** (Tokens/s)
Secondary KPI	Variance Reduction (Stability)
Deliverable	Auditable Evidence Artifacts (summary.json)

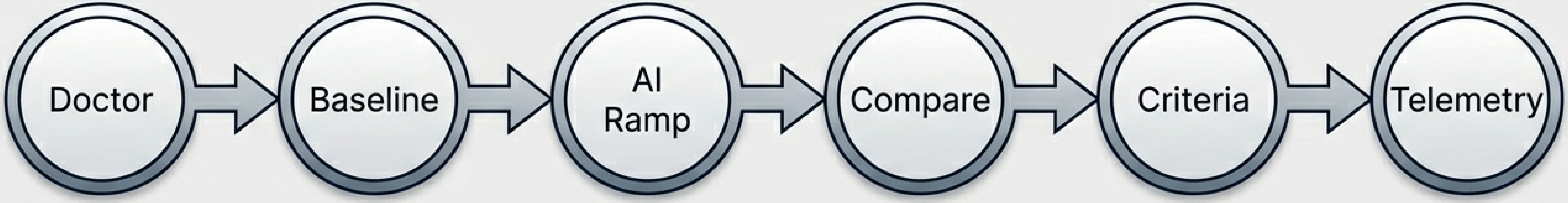
Reference Architecture



Scale-Up (Single Node)
High Core-Count CPU & RAM
Context: Tensor Parallelism in vLLM splits model weights across GPUs.
8x GPUs
(AMD ROCm / NVIDIA CUDA)
NVMe Scratch Storage
Model Cache
Scale-Out (Rack Scale)
Scheduler: Slurm / MPI / K8s
N Nodes
(8x GPU each)
Shared Filesystem / Artifact Aggregation

The Golden Path Protocol

Deterministic Optimization Loop



Determinism

Baseline and AI Ramp runs are identical except for the optimization layer.

Auditability

RUN_ID-based outputs prevent pathing errors. The ‘After’ state is strictly comparable to the ‘Before’ state.

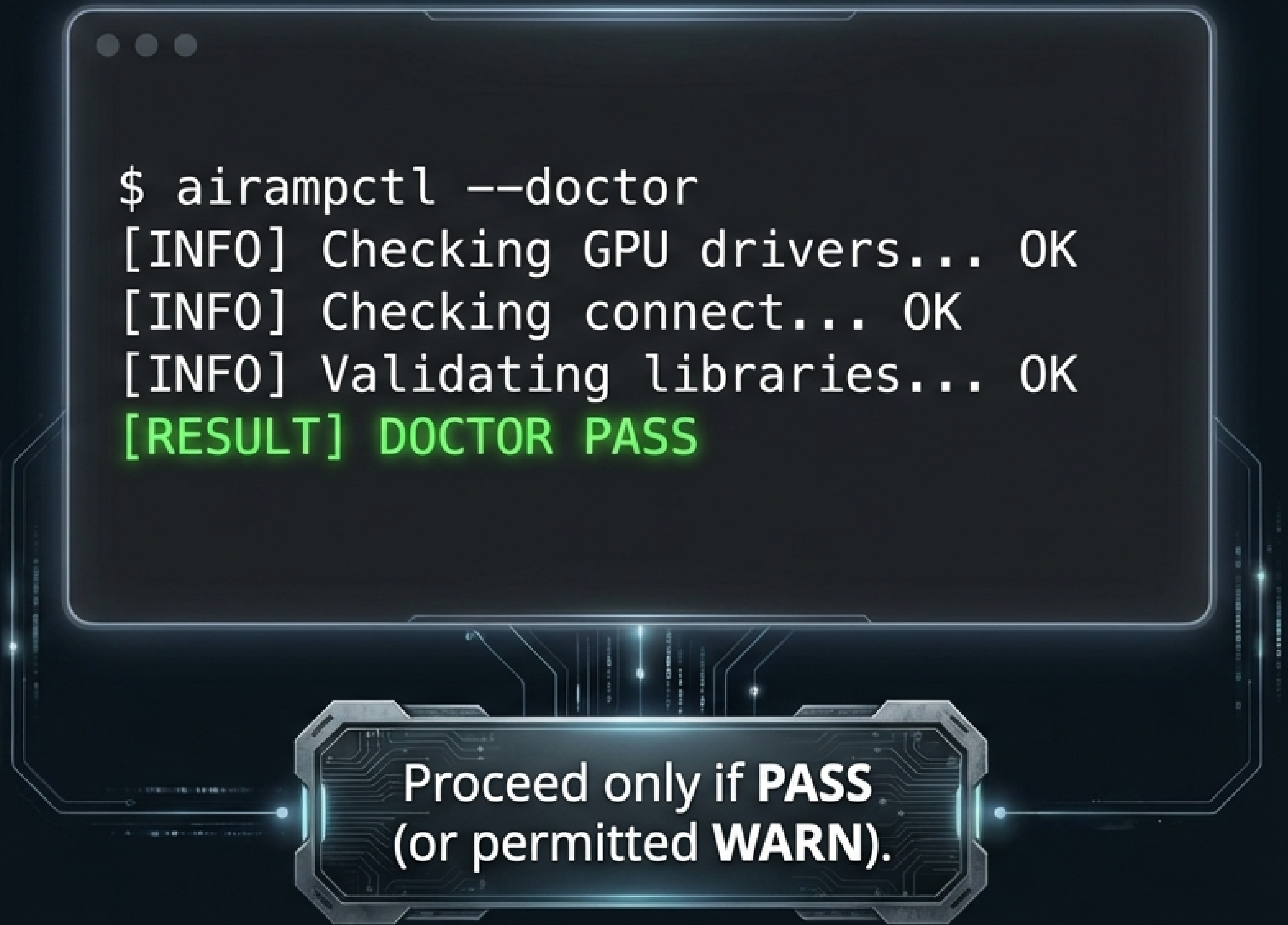
Step 0 & 1: Validation

Step 0: Installation

```
./install.sh --runtime
./install.sh --bench
```

Step 1: The Doctor

Validates drivers, connectivity, and libraries.



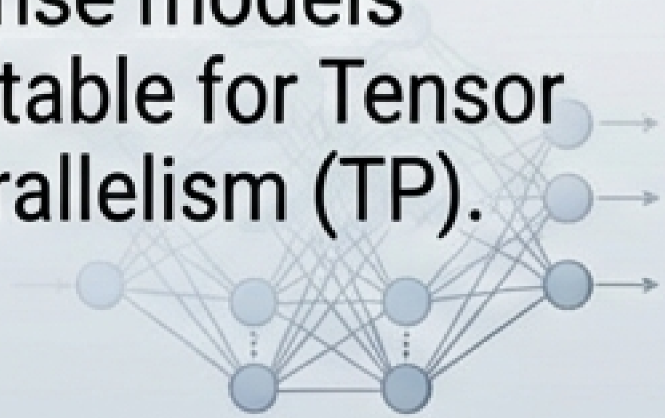
```
$ airampctl --doctor
[INFO] Checking GPU drivers... OK
[INFO] Checking connect... OK
[INFO] Validating libraries... OK
[RESULT] DOCTOR PASS
```

Proceed only if **PASS** (or permitted **WARN**).

Target Model Classes

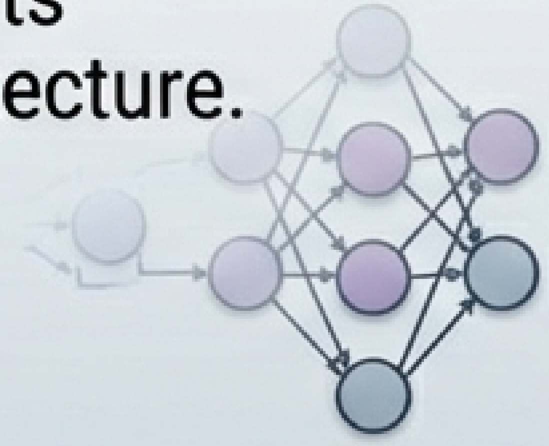
Ungated 70B-Class Dense

- e.g., Qwen2-72B / Qwen2.5-72B.
- Configuration: Dense models suitable for Tensor Parallelism (TP).



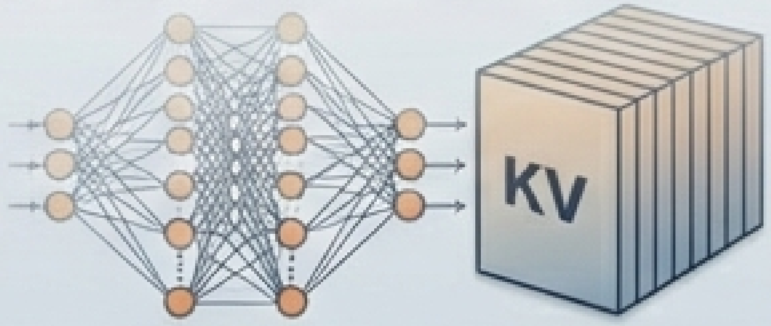
High-Quality MoE

- e.g., Mixtral-8x22B.
- Configuration: Sparse Mixture of Experts architecture.



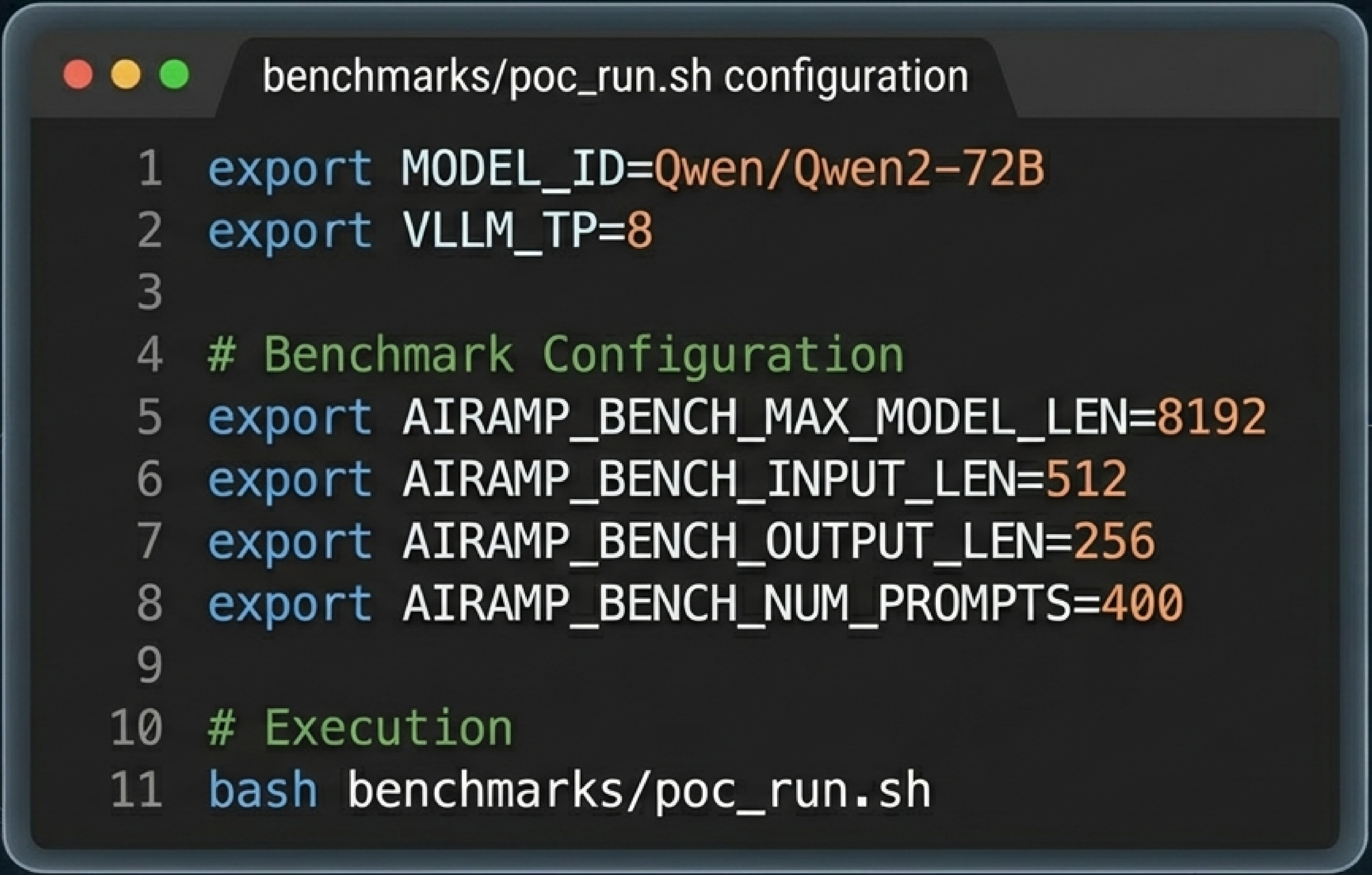
700B-Class MoE

- e.g., DeepSeek-V3.
- Scale: ~671B Total Params, ~37B Active. Requires massive KV cache headroom.



Licensing Context: 'Ungated' implies general accessibility (HuggingFace), but commercial terms (e.g., Qwen License) still apply.

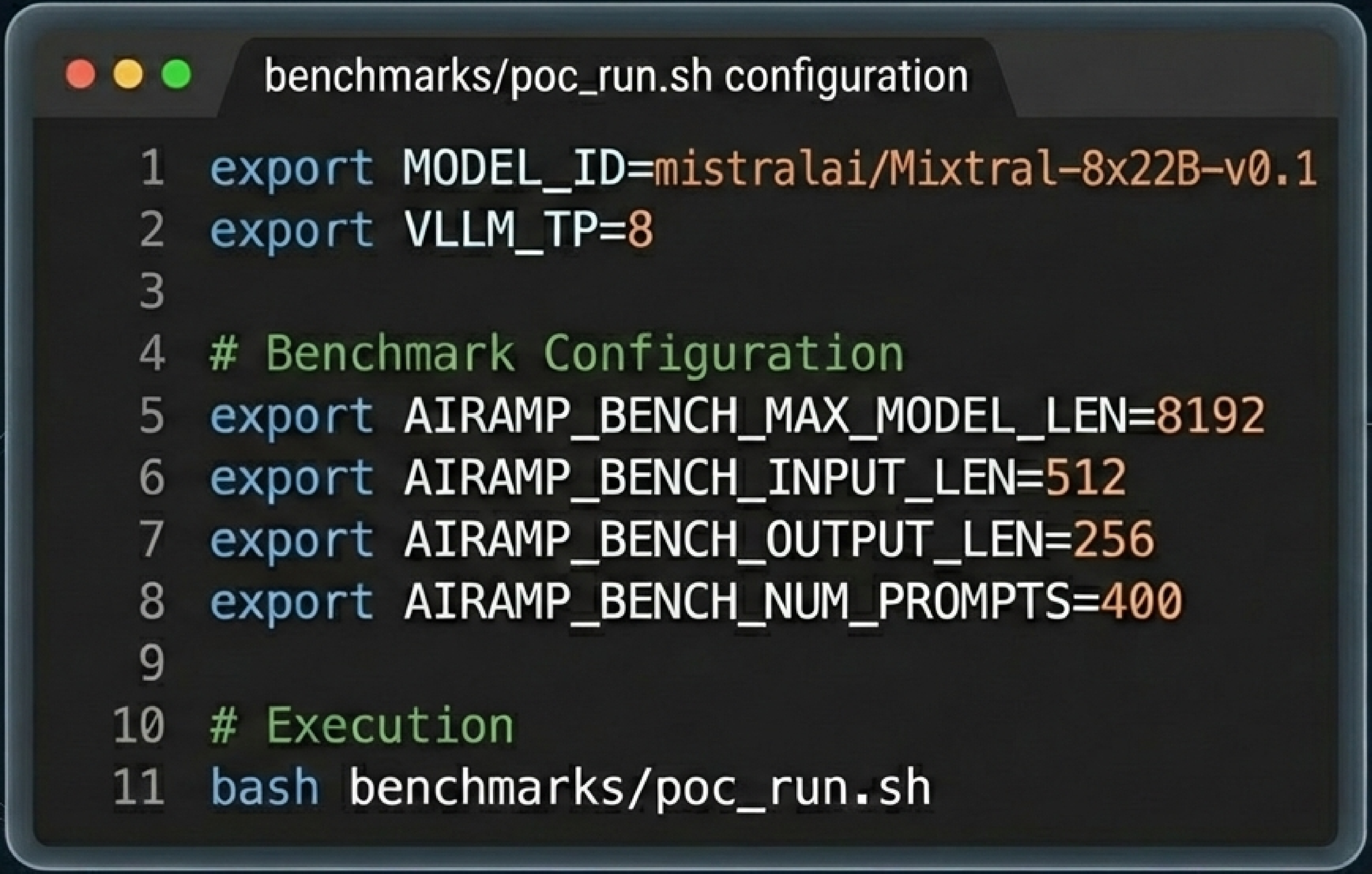
Implementation Recipe: Qwen2-72B (Dense TP)



```
export MODEL_ID=Qwen/Qwen2-72B
export VLLM_TP=8

# Benchmark Configuration
export AIRAMP_BENCH_MAX_MODEL_LEN=8192
export AIRAMP_BENCH_INPUT_LEN=512
export AIRAMP_BENCH_OUTPUT_LEN=256
export AIRAMP_BENCH_NUM_PROMPTS=400

# Execution
bash benchmarks/poc_run.sh
```

Standard **8-GPU** Tensor Parallel setup.

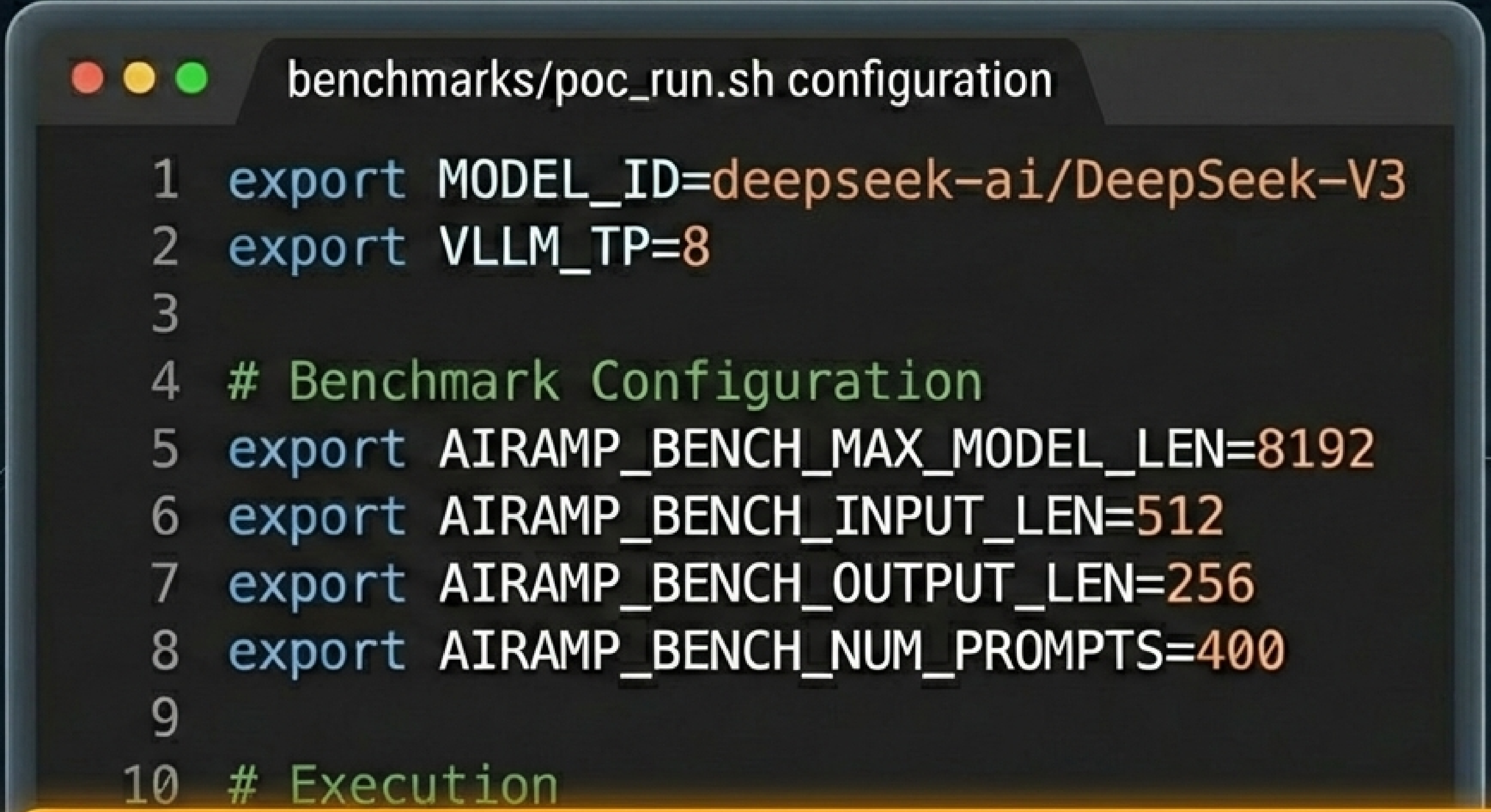
Implementation Recipe: Mixtral-8x22B (MoE)



benchmarks/poc_run.sh configuration

```bash
export MODEL_ID=mistralai/Mixtral-8x22B-v0.1
export VLLM_TP=8

# Benchmark Configuration
export AIRAMP_BENCH_MAX_MODEL_LEN=8192
export AIRAMP_BENCH_INPUT_LEN=512
export AIRAMP_BENCH_OUTPUT_LEN=256
export AIRAMP_BENCH_NUM_PROMPTS=400

# Execution
bash benchmarks/poc_run.sh
```

Implementation Recipe: DeepSeek-V3 (700B Class)



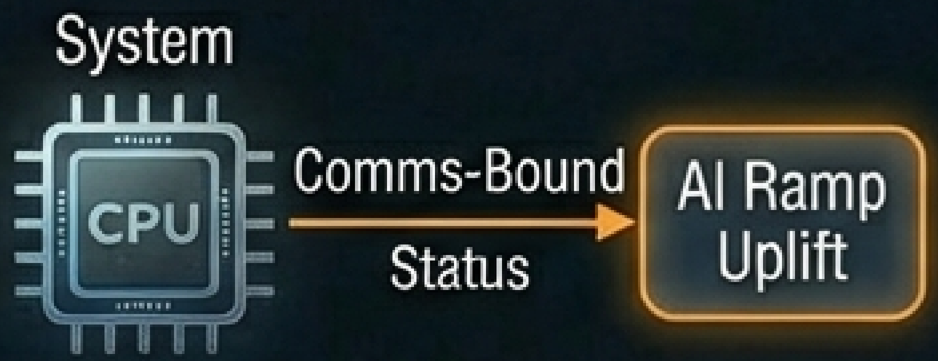
```
export MODEL_ID=deepseek-ai/DeepSeek-V3
export VLLM_TP=8

# Benchmark Configuration
export AIRAMP_BENCH_MAX_MODEL_LEN=8192
export AIRAMP_BENCH_INPUT_LEN=512
export AIRAMP_BENCH_OUTPUT_LEN=256
export AIRAMP_BENCH_NUM_PROMPTS=400

# Execution
```

STORAGE WARNING: DeepSeek-V3 is a massive MoE checkpoint. Ensure adequate NVMe scratch space and GPU capacity before execution.

Tuning for Uplift: The Comms-Bound Playbook

Targeting stress conditions for maximum optimization.

The Logic: Tensor parallel inference is limited by cross-GPU sync, collective ops, and KV cache pressure. AI Ramp provides the most uplift when the system is “Comms-Bound”.


System

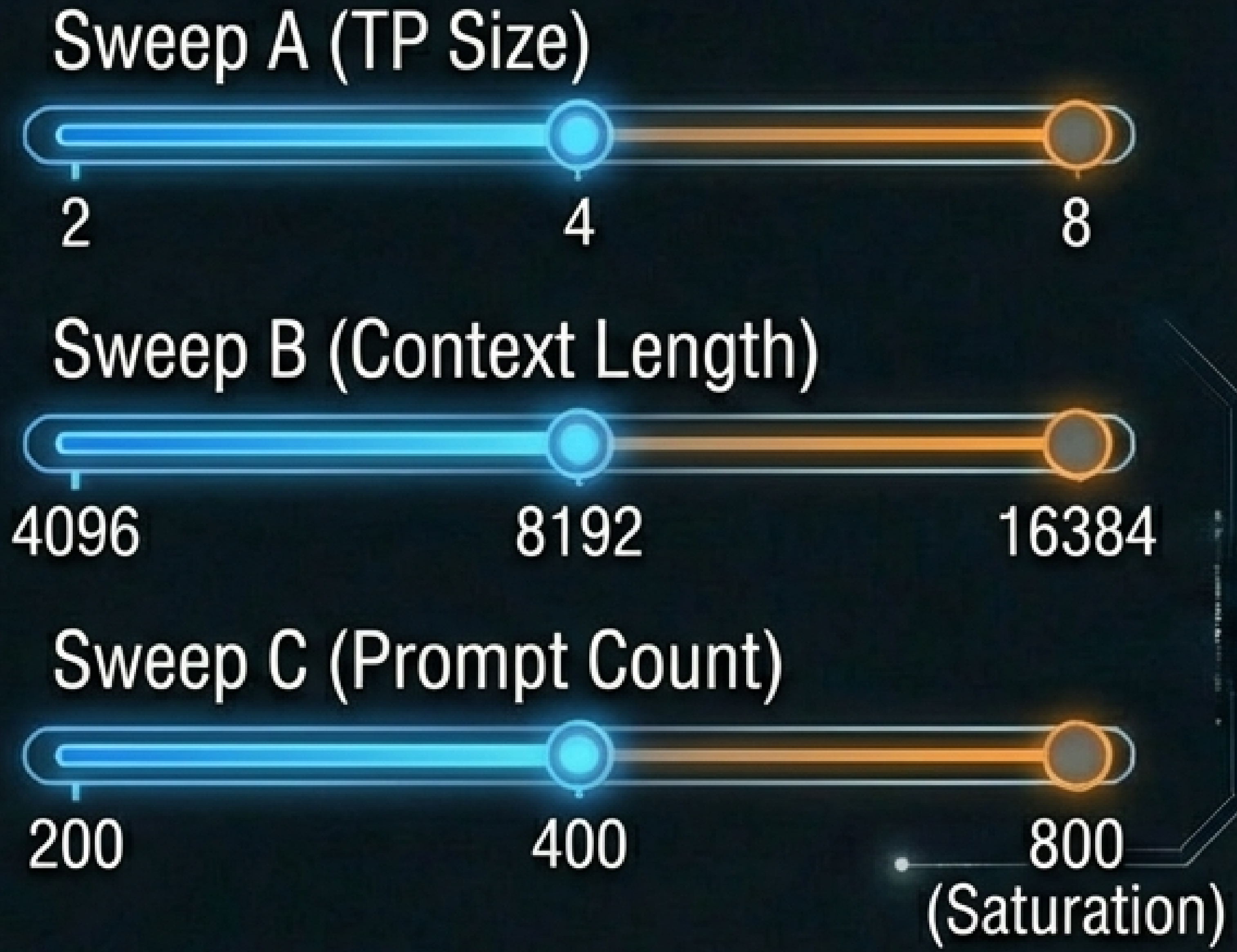
CPU

Comms-Bound Status

AI Ramp Uplift

```
Enable Stress Mode:
export AIRAMP_BENCH_REQUIRE_COMMS_BOUND=1
```

Parameter Sweeps


Sweep A (TP Size)

2 4 8

Sweep B (Context Length)

4096 8192 16384

Sweep C (Prompt Count)

200 400 800 (Saturation)

Scaling Out: Rack Scale Execution

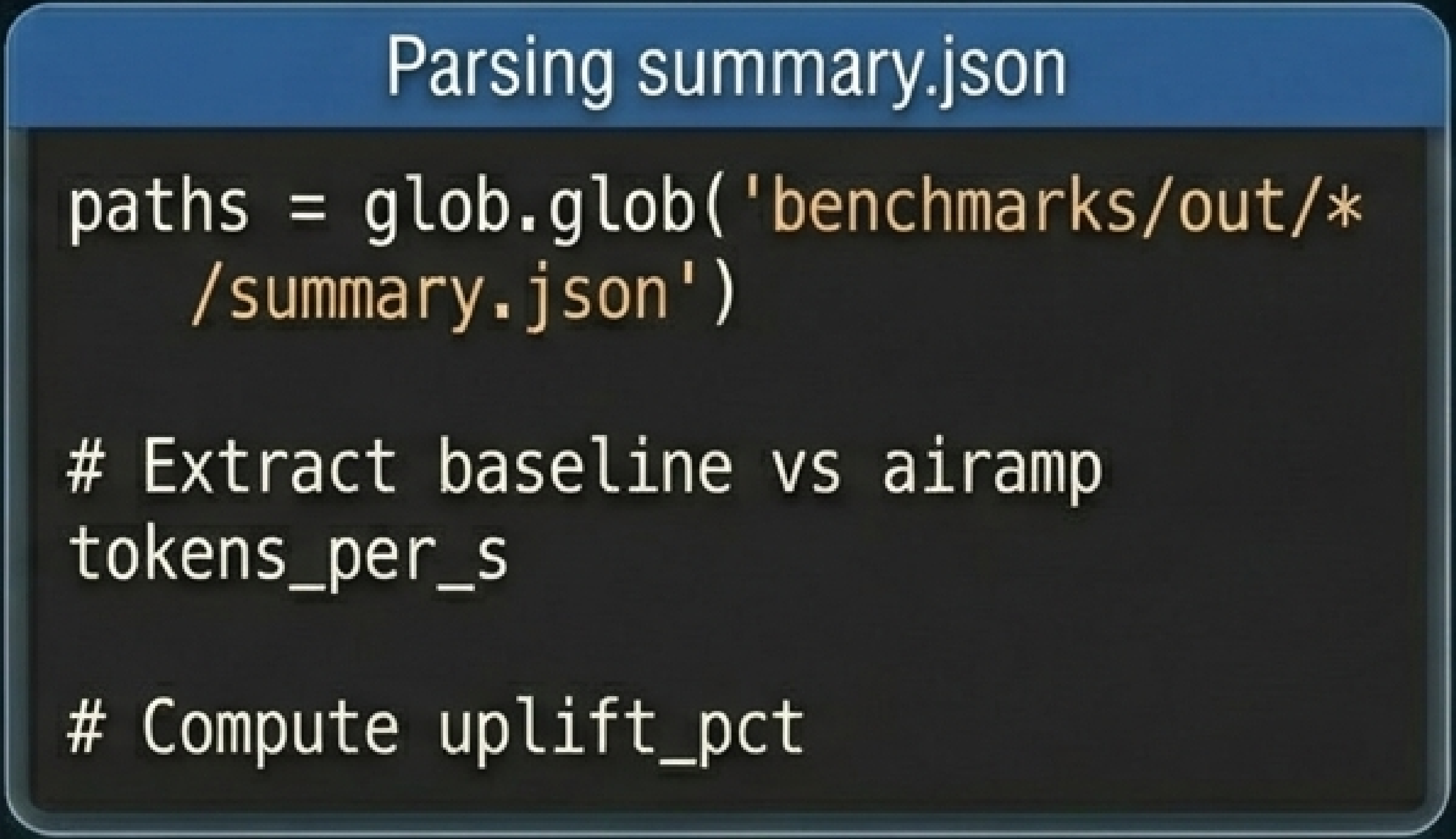
Slurm Job Template

```
#SBATCH -N 1
#SBATCH --gres=gpu:8
#SBATCH --exclusive

# Unique Tracking
RUN_ID=$(date -u +%Y%m%dT%H%M%SZ)
_${SLURM_JOB_ID}_$(hostname)
```

```
$ sbatch -N 1 --array=1-8 slurm_airamp_poc.sh
```

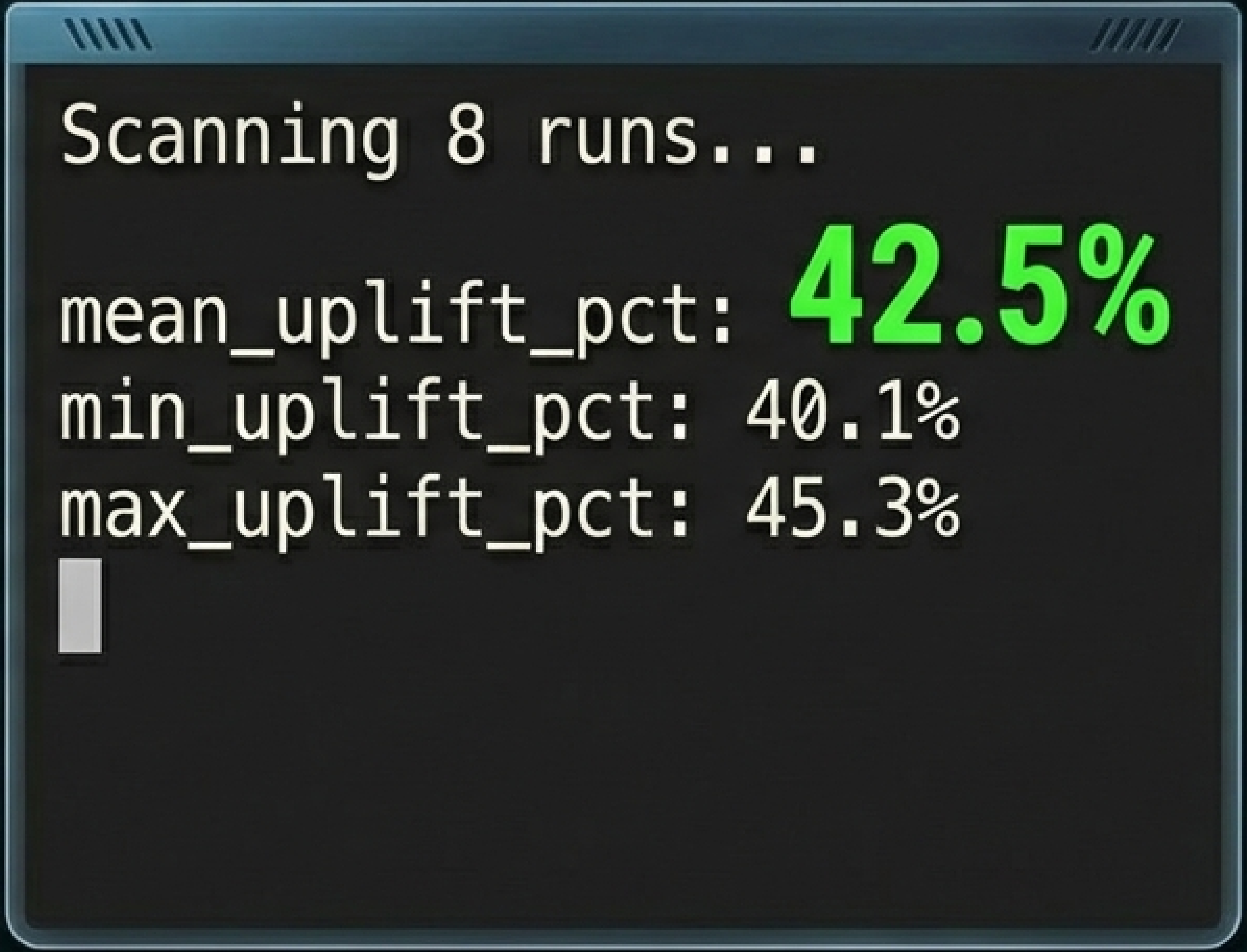
8 Simultaneous Node Executions

Results Aggregation



Parsing summary.json

```
paths = glob.glob('benchmarks/out/*
    /summary.json')

# Extract baseline vs airamp
tokens_per_s

# Compute uplift_pct
```



```
Scanning 8 runs...

mean_uplift_pct: 42.5%
min_uplift_pct: 40.1%
max_uplift_pct: 45.3%
█
```

DOE / Utility Addendum: Energy & Carbon

$$Tokens\ per\ Joule = \frac{Tokens\ per\ Second}{Watts}$$

Measuring Power

NVIDIA

```
nvidia-smi --query-gpu=power.draw ...
```

AMD

```
rocm-smi --showpower
```

Grid Impact

- Higher throughput under fixed power envelopes.
- Deferred capacity expansion.
- Reduced carbon intensity per inference.

Verification & Reporting

Evidence Artifacts

```
benchmarks/out/<RUN_ID>/summary.json
```

`compare_output.txt` (Executive Delta)

```
support_bundle.tar.gz
```

Acceptance Criteria

✓ Throughput uplift ≥40%

✓ Doctor returns PASS

✓ Compare exits with code 0

Support Command

```
bash benchmarks/make_support_bundle.sh
<RUN_ID>
```

Summary & Troubleshooting

Operator Troubleshooting

Doctor FAIL?

-> Stop. Remediate drivers/paths.

Compare Errors?

-> Verify RUN_ID matches.

Wrong Paths?

-> v196 auto-recovers unambiguous paths.

Final Takeaways

- AI Ramp delivers ≥40% uplift via software runtime optimization.
- The ‘Golden Path’ ensures deterministic, auditable results.
- Validated for Qwen, Mixtral, and DeepSeek on 8-GPU nodes.

TENSOR™ NETWORKS

AI Ramp