

3D Parallel Execution Engine

KVP-0

SPP-0



SPP-1



KVP-1

SPP-0



SPP-1



Replica Controller

Batch Scheduler

Batch Packer

Runtime Predictor

Prefill Priority Queue

1 Add New Request

4 Get Least Slack Prefill

3 Update Batch

4 Check Runtime Constraint

7 Return Generated Tokens

5 Dispatch Batch

6 Reque Partial Prefill

Client Interface

Request

Response

