Optimization comes at different levels ...

cluster optimization

- cluster-wide, constrained cost minimization
- heterogeneous GPUs (re-)allocation, dynamic
- multi-server scaling tradeoffs, cluster scaling, bursting

queueing optimization

- request queueing, request SLOs
- request routing
- single server scaling

inference server optimization

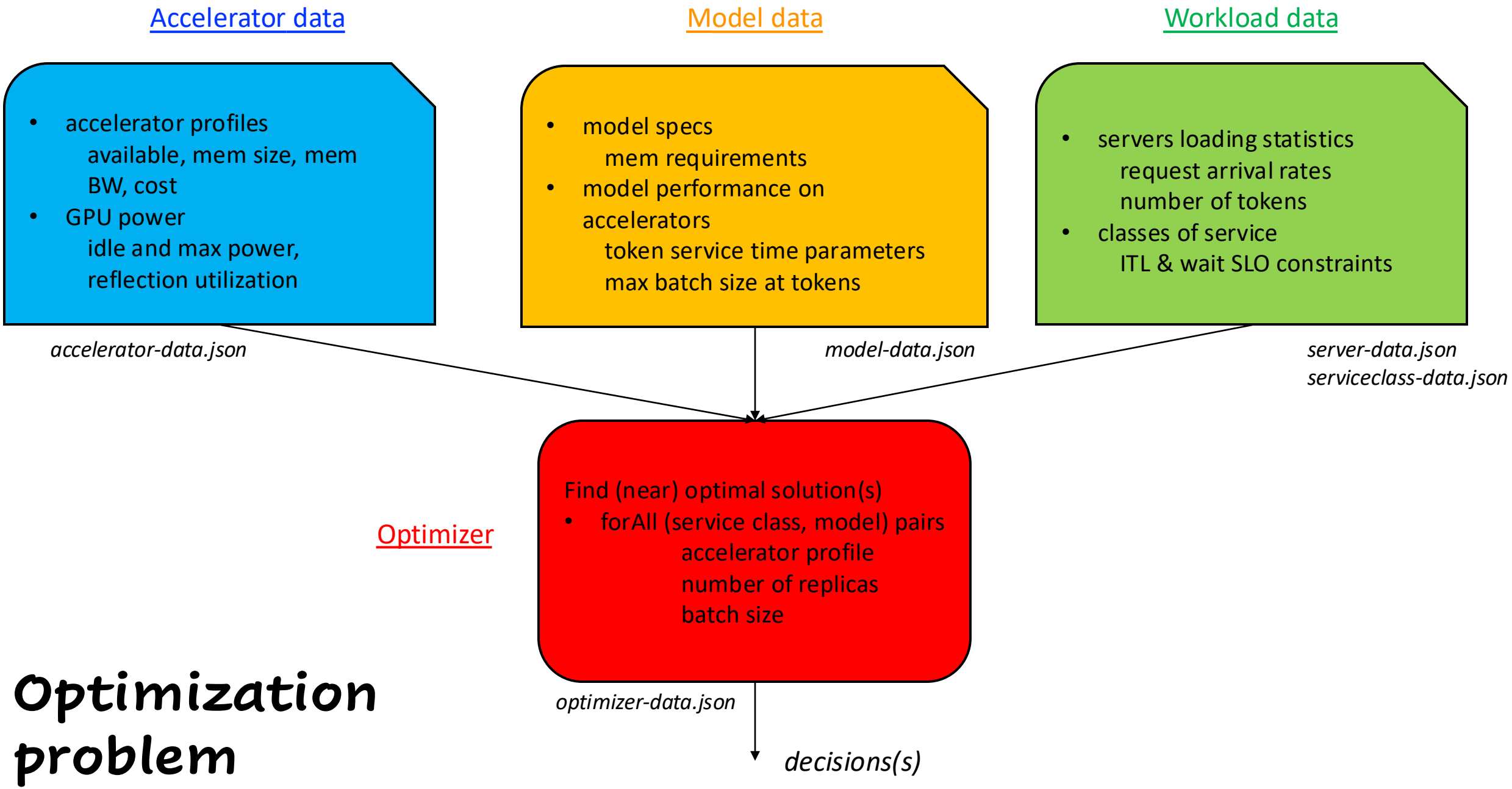
- parallelism, K/V cache management
- prefill/decode scheduling
- caching/swapping, adapters management

model optimization

- memory, compute efficiency, design
- accuracy, quantization, tuning

device optimization

- driver, allocation, DRA
- SM scheduler, multiplexing, sharing



Optimization problem

accelerator-data.json

```
"MI300X": {
    "type": "MI300X",
    "multiplicity": 1,
    "memSize": 192,
    "memBW": 5300,
    "power" : {
        "idle": 220,
        "full": 750,
        "midPower": 650,
        "midUtil": 0.6
    },
    "cost": 65.00
},
"2xA100": {
    "type": "A100",
    "multiplicity": 2,
    "memSize": 160,
    "memBW": 4000,
    "power" : {
        "idle": 300,
        "full": 800,
        "midPower": 640,
        "midUtil": 0.6
    },
    "cost": 80.00
},
```

model-data.json

```
{
  "name": "granite_34b",
  "acc": "H100",
  "accCount": 2,
  "alpha": 20.49,
  "beta": 0.34,
  "maxBatchSize": 12,
  "atTokens": 512
},
{
  "name": "llama3_8b",
  "acc": "MI300X",
  "accCount": 1,
  "alpha": 4.88,
  "beta": 0.22,
  "maxBatchSize": 38,
  "atTokens": 512
},
```

serviceclass-data.json

```
{
    "name": "Premium",
    "model": "llama_7b",
    "slo-itl": 40,
    "slo-ttw": 500
},
{
    "name": "Bronze",
    "model": "llama_7b",
    "slo-itl": 80,
    "slo-ttw": 1000
},
{
    "name": "Free",
    "model": "mistral_7b",
    "slo-itl": 160,
    "slo-ttw": 2000
},
{
    "name": "Batch1K",
    "model": "llama_13b",
    "priority": 3,
    "slo-tps": 1000
},
```

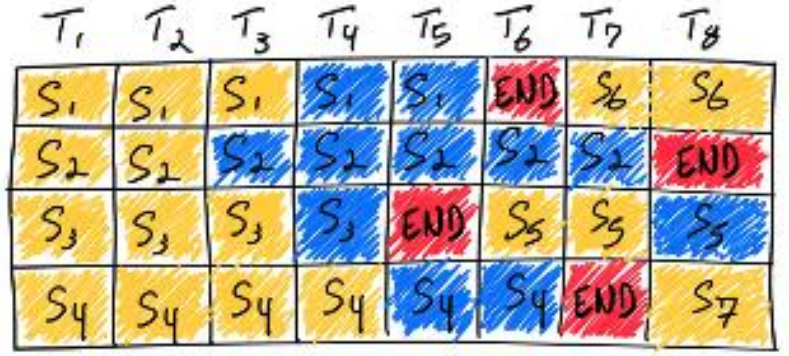
server-data.json

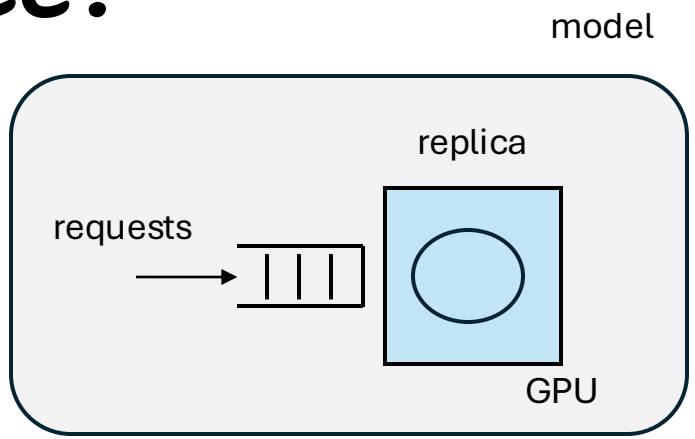
```
{
    "name": "Free-llama3_8b",
    "class": "Free",
    "model": "llama3_8b",
    "currentAlloc": {
        "accelerator": "MI250",
        "numReplicas": 12,
        "maxBatch": 12,
        "cost": 552,
        "itlAverage": 11.231181,
        "waitAverage": 558.5869,
        "load": {
            "arrivalRate": 480,
            "avgLength": 1024,
            "arrivalCOV": 1.5,
            "serviceCOV": 1.5
        }
    },
    "desiredAlloc": {
        "accelerator": "MI300X",
        "numReplicas": 5,
        "maxBatch": 19,
        "cost": 325,
        "itlAverage": 7.924246,
        "waitAverage": 325.02734,
        "load": {
            "arrivalRate": 480,
            "avgLength": 1024,
            "arrivalCOV": 1.5,
            "serviceCOV": 1.5
        }
    }
}
```

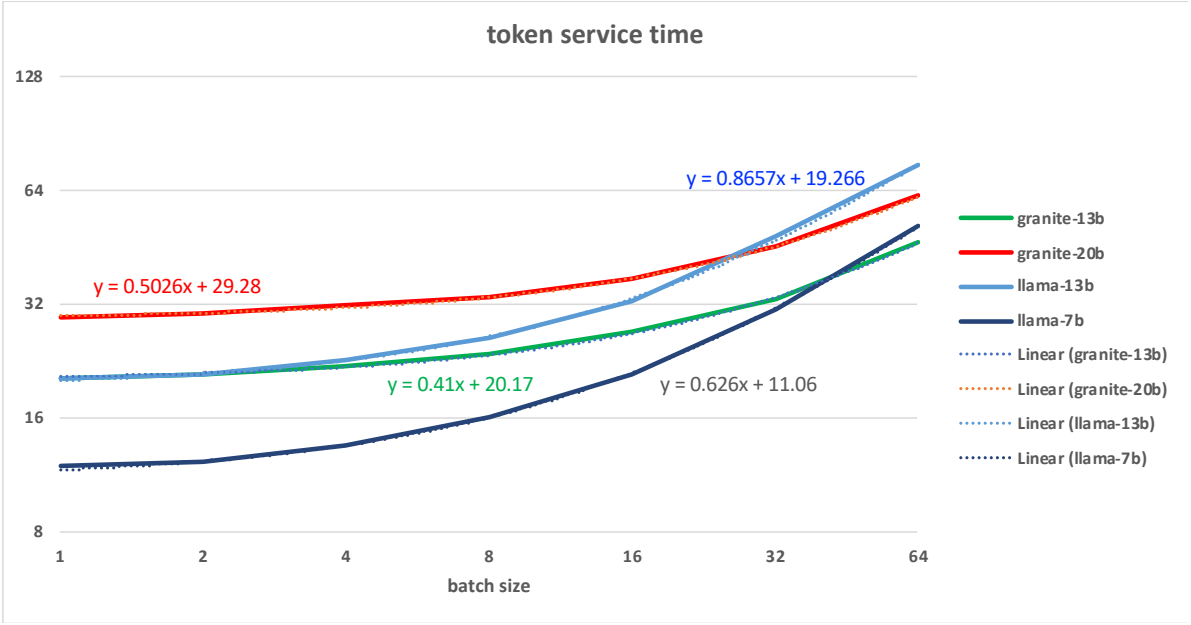
input

output

how to model batching of request service?







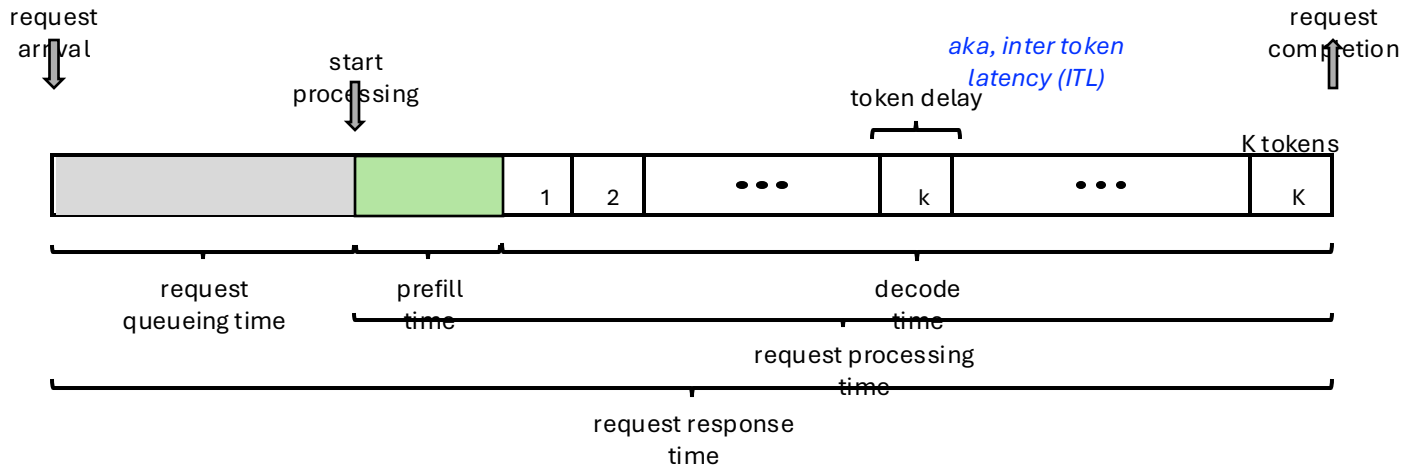
Fitting token service time based on benchmarking of pairs of models and accelerators

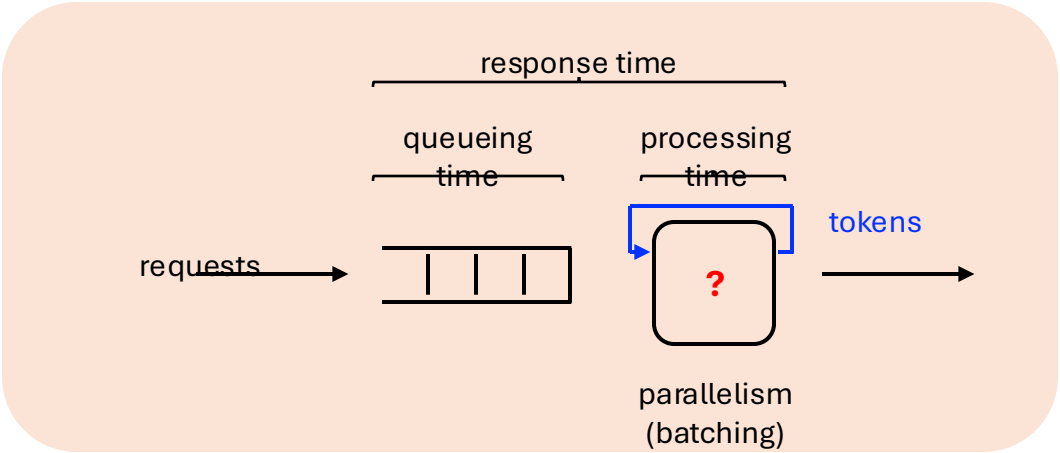
$$T(n) \approx \alpha + \beta n$$

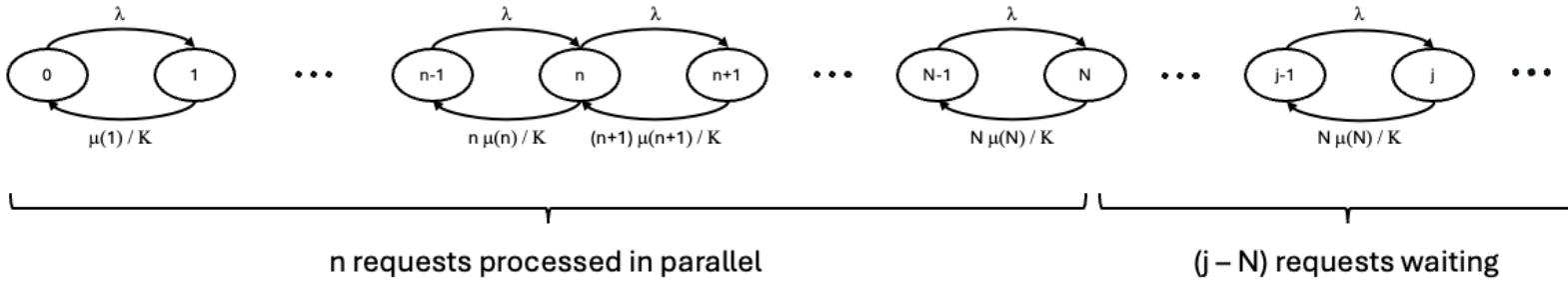
queueing modeling

- Markovian assumptions

- state: number of requests in process
- λ request arrival rate
- K average number of tokens per request
- $\tau(n)$ average token service time given n batches
- N batch size (maximum)







$$\mu(n) = \frac{1}{T(n)}$$

$$T(n) \approx \alpha + \beta\, n$$

modeling batching

Define

$$g(n) = \frac{\lambda\, K}{n\, \mu(n)}, \quad n = 1, 2, \cdots, N.$$

Let π_n, $n \geq 0$, be the steady state probability that there are n requests in the system. Using the birth-death chain, we get

$$\pi_n = \begin{cases} \pi_0 \prod_{i=1}^{n} g(i), & n = 1, 2, \cdots, N, \\ \pi_N\, g(N)^{(n-N)}, & n > N, \end{cases}$$

where

$$\pi_0 = 1 - \sum_{n=1}^{\infty} \pi_n, \quad \pi_0 > 0.$$

modeling batching

Given

$$\tau(n) = \begin{cases} \alpha + \beta n, & n = 1, 2, \cdots, N, \\ \alpha + \beta N, & n > N, \end{cases}$$

and

$$g(n) = \frac{\lambda\, K}{n\, \mu(n)}, \quad n = 1, 2, \cdots, N,$$

we get the special cases

$$g(n) = \begin{cases} \lambda \beta K, & \alpha = 0, \\ \lambda \alpha K / n, & \beta = 0, \end{cases}$$

$n = 1, 2, \cdots, N$, which correspond to single and multiple server queues, respectively.

ITL limiting using queueing model

requestRate	35.200	/min	p(0)	1.0542E-08		avgNumSystem	30.23
lambda	0.00059	/msec	maxN	200		avgNumServer	30.04
K	1024		pFull	1.4442E-14		avgNumQueue	0.20
N	48		effecTput	35.20031	req/m	utilization	0.626
alpha	19			0.00059	/msec	avgRespTime	51,533
beta	1		P(N)	0.96847		avgSrvTime	51,200
lambda*K	0.600752					avgWaitTime	333
						avgTokenServTime	50.0

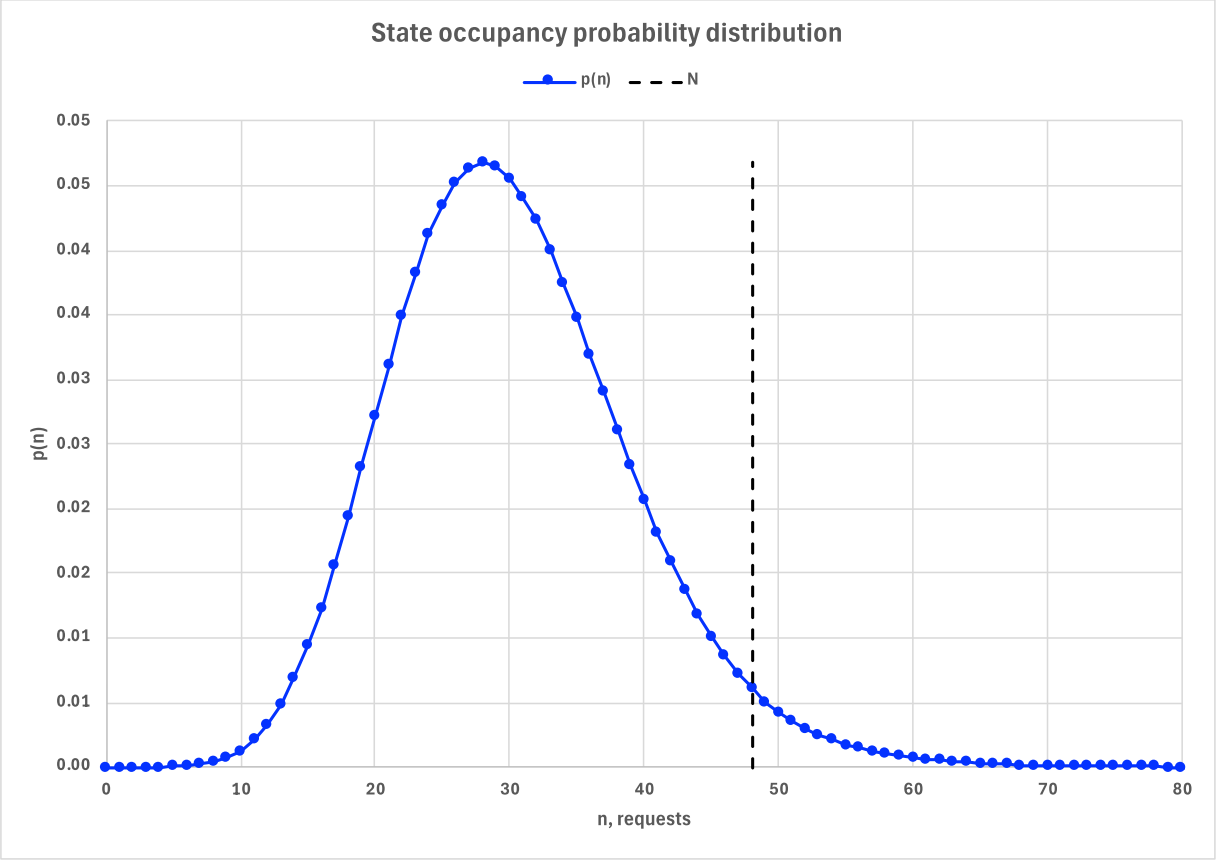
Find

maximum request rate

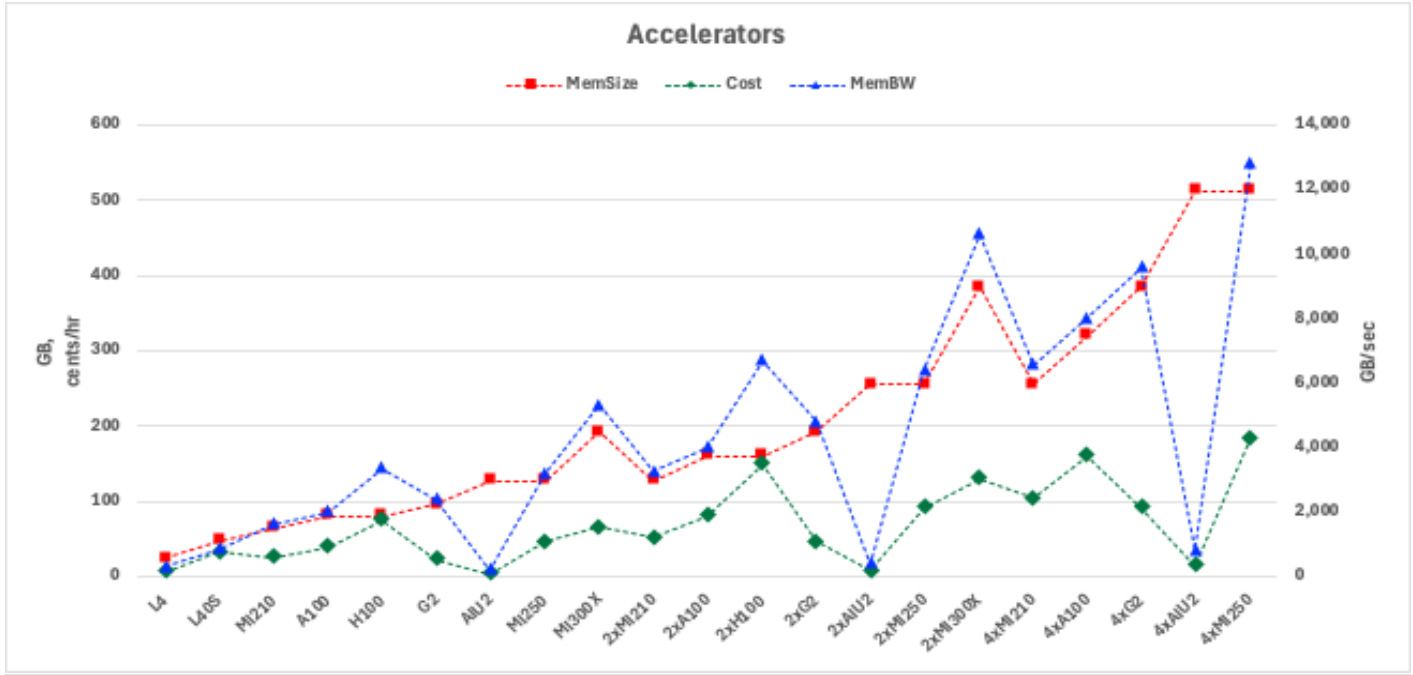
s.t.

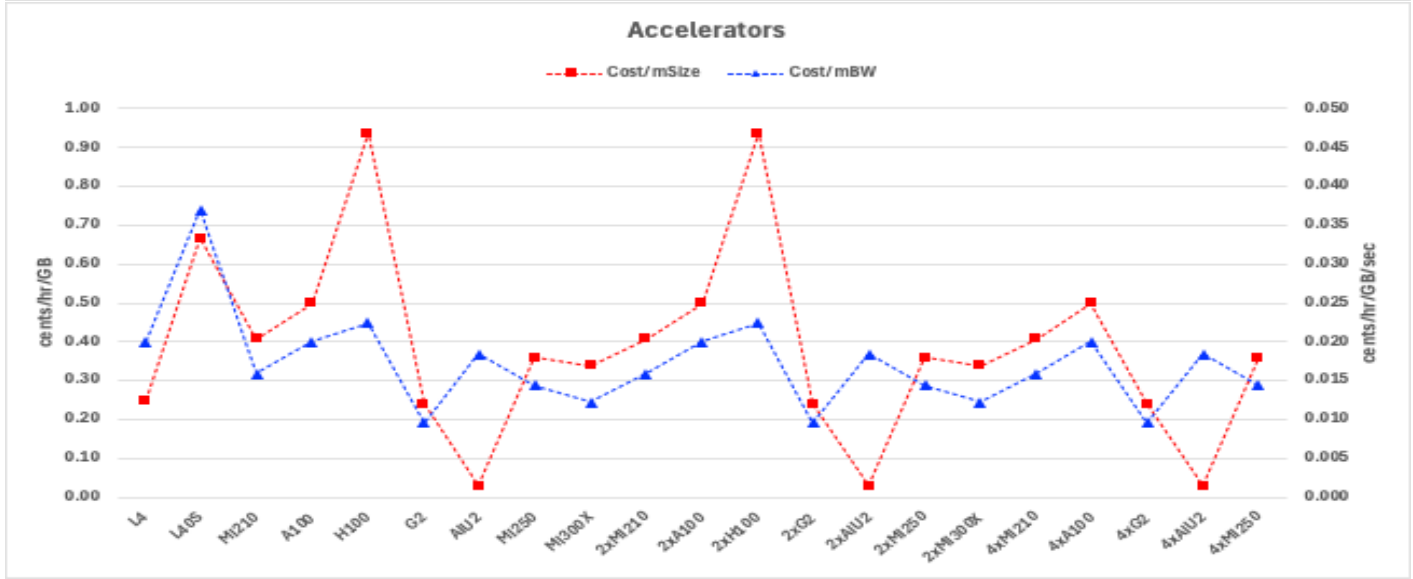
avgTokenServTime <= target

target = 50 msec
requestRate = 35.2 req/min

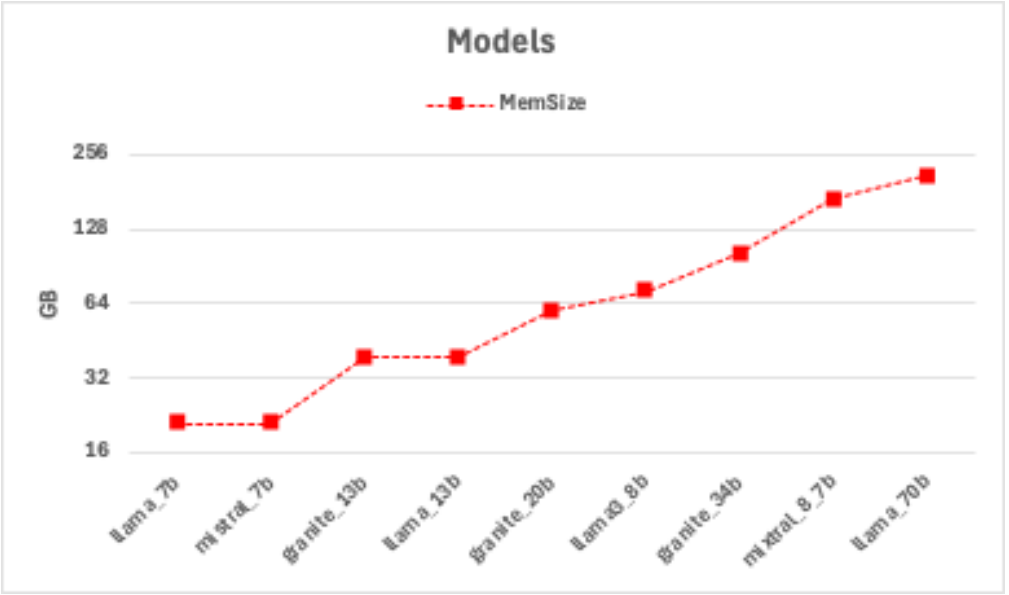


Accelerators Specs



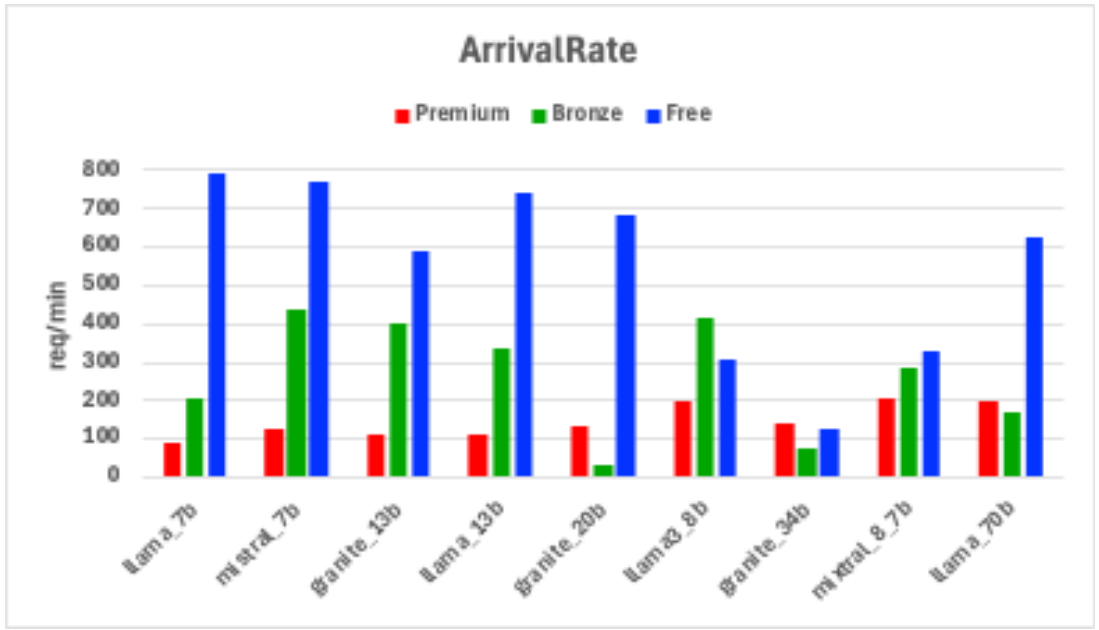


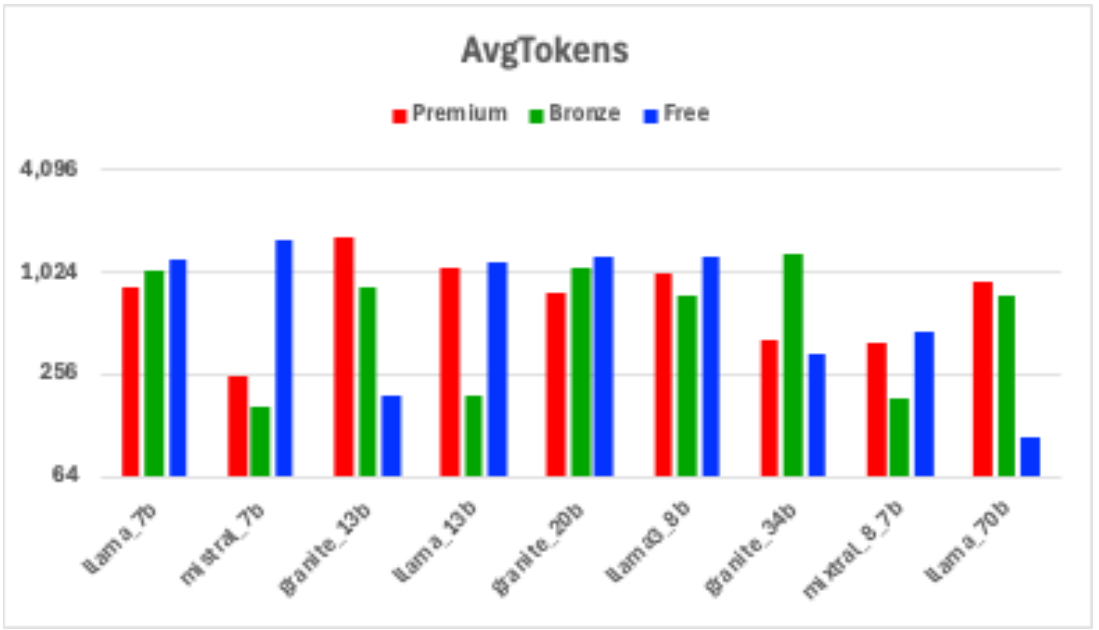
Models Specs



Unlimited accelerators - Dynamic

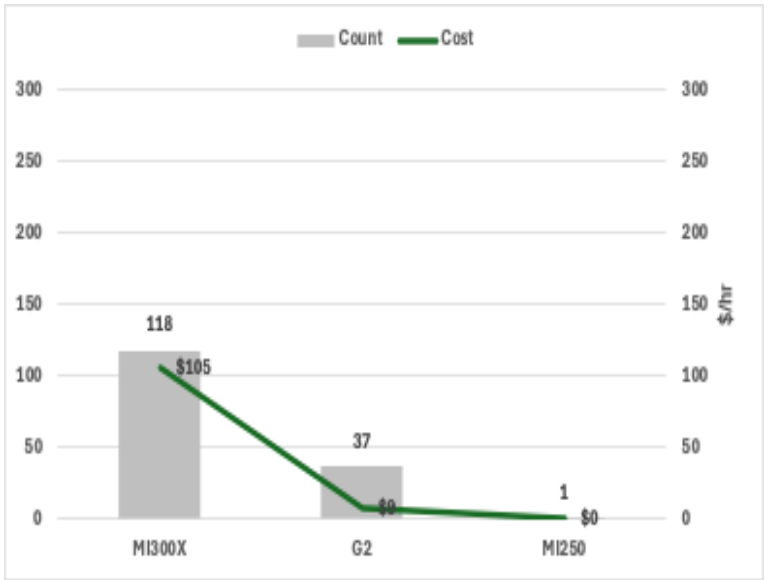
- change in request rates
- change in request lengths
- change/scale accelerators
- minimize change in accelerators and cost





Accelerator Change

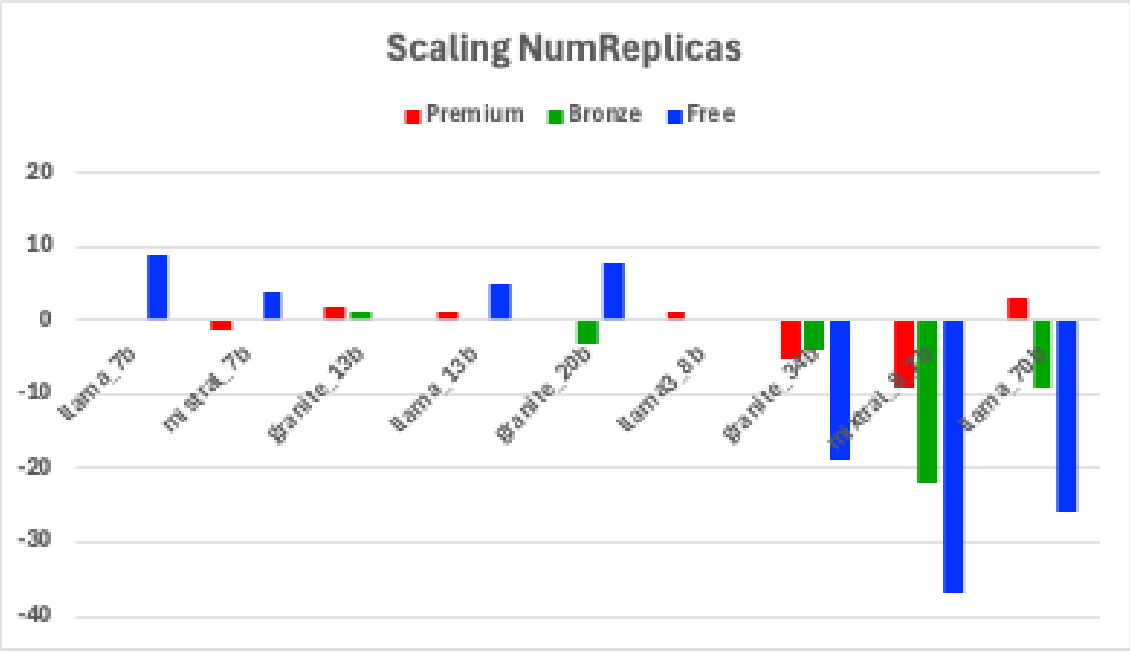
	Premium	Bronze	Free
llama_7b			
mistral_7b		MI300X->G2	
granite_13b			MI300X->G2
llama_13b		MI300X->MI250	
granite_20b	MI300X->G2		
llama3_8b			
granite_34b			
mixtral_8_7b			
llama_70b			

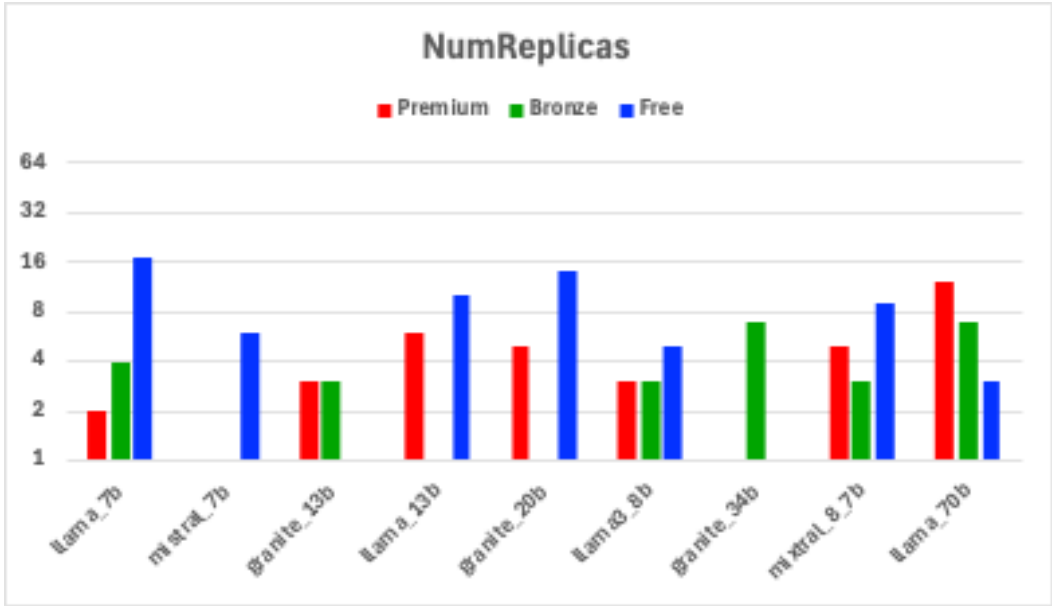


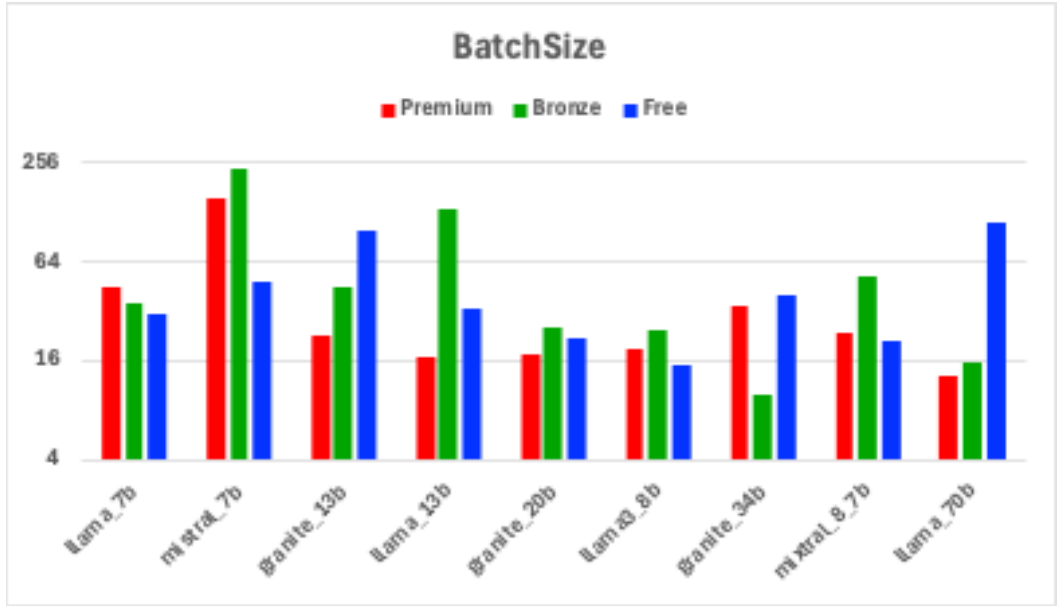
TotalCost
11,427.00

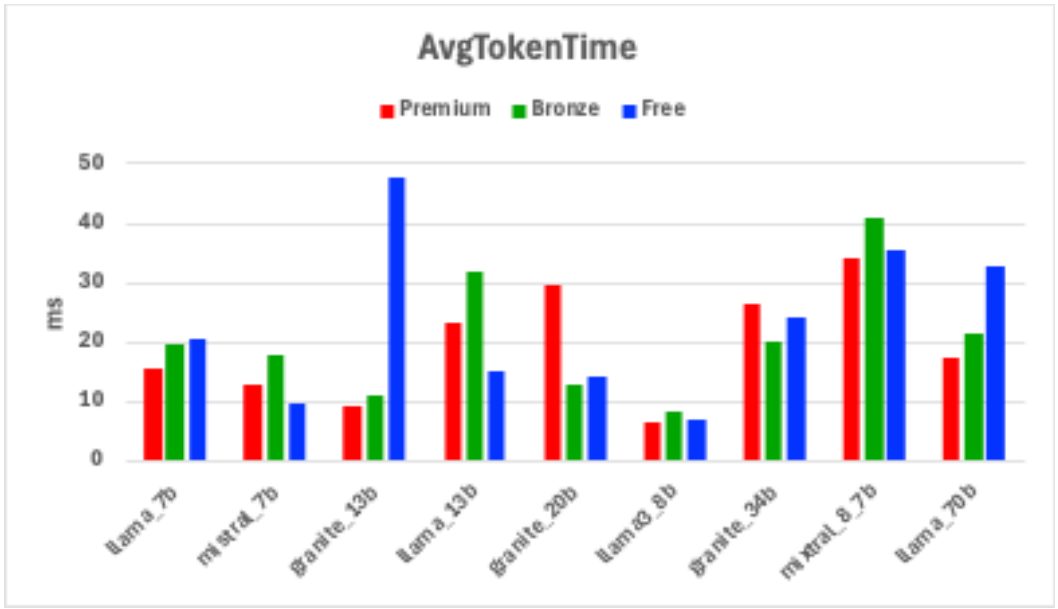
Accelerator Change			
	Premium	Bronze	Free
llama_7b			
mistral_7b		MI300X->G2	
granite_13b			MI300X->G2
llama_13b	MI300X->G2	MI300X->MI250	
granite_20b			
llama3_8b			
granite_34b			
mixtral_8_7b			
llama_70b			

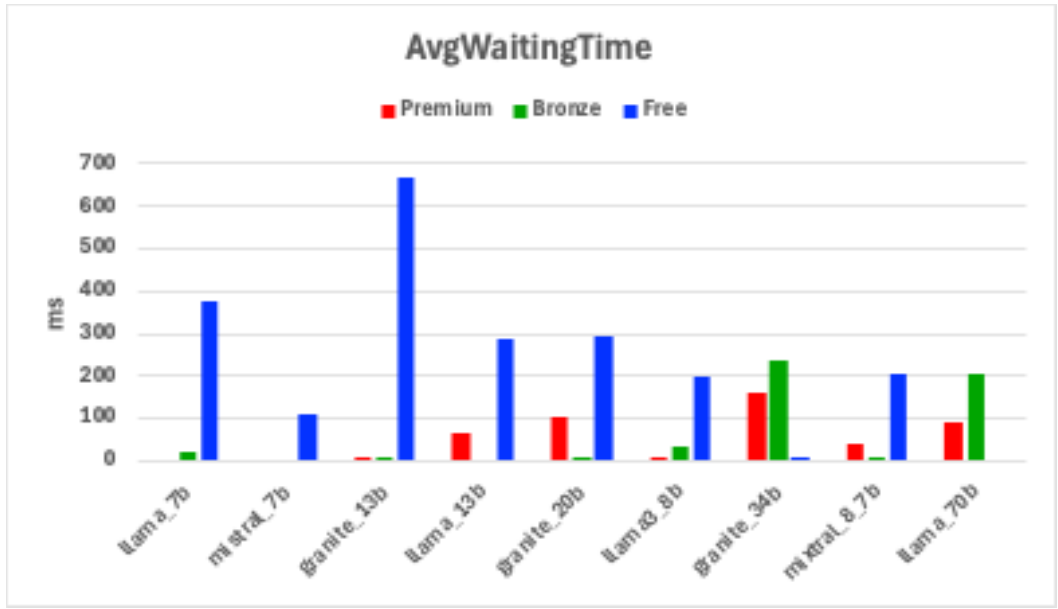
Scale			
	Premium	Bronze	Free
llama_7b	0	0	9
mistral_7b	-1		4
granite_13b	2	1	
llama_13b	1		5
granite_20b		-3	8
llama3_8b	1	0	0
granite_34b	-5	-4	-19
mixtral_8_7b	-9	-22	-37
llama_70b	3	-9	-26

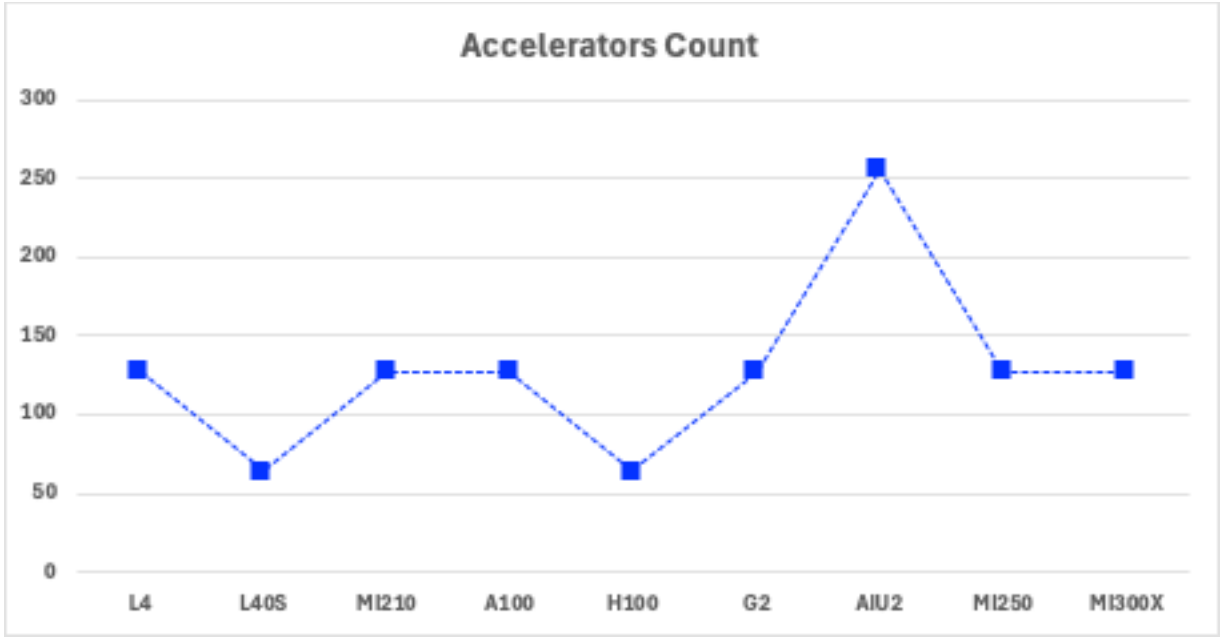



NumReplicas
Premium
Bronze
Free
64
32
16
8
4
2
1
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b


BatchSize
Premium
Bronze
Free
256
64
16
4
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b

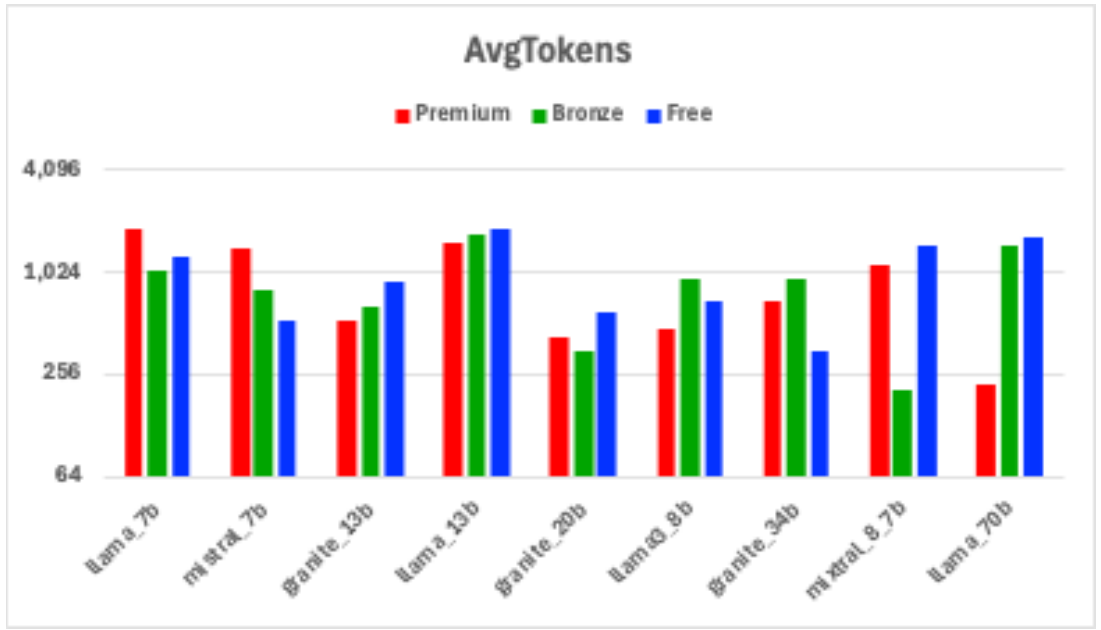

AvgTokenTime
Premium
Bronze
Free
ms
50
40
30
20
10
0
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b


AvgWaitingTime
Premium
Bronze
Free
ms
700
600
500
400
300
200
100
0
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b

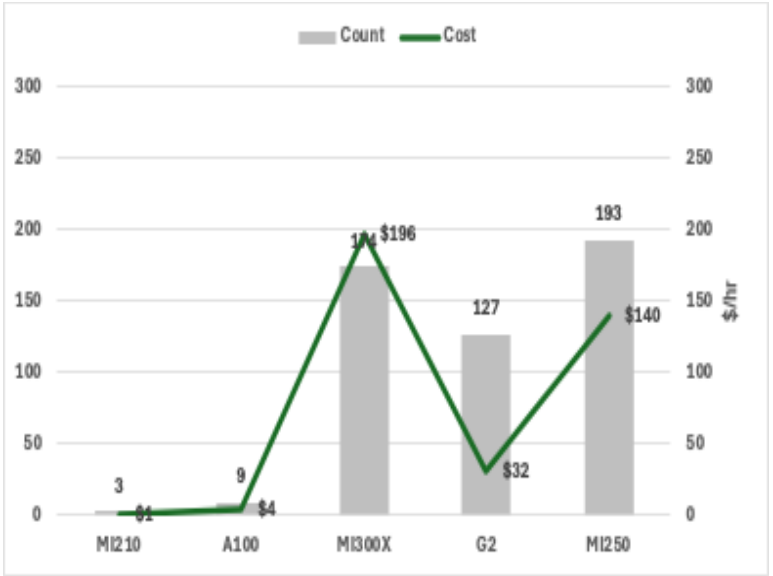


Limited accelerators - Dynamic

- change in request rates
- change in request lengths
- change/scale accelerators
- minimize change in accelerators and cost



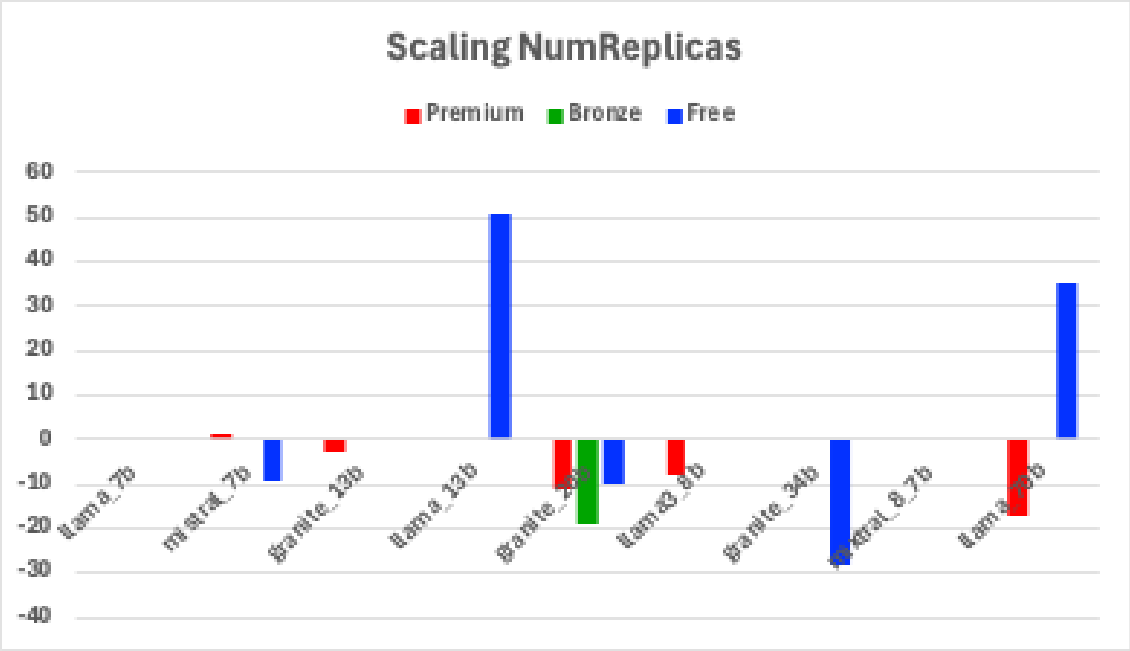
Accelerator Change			
	Premium	Bronze	Free
llama_7b	MI210->MI250	MI210->G2	MI210->MI250
mistral_7b		MI210->MI250	
granite_13b		MI300X->MI250	A100->G2
llama_13b	A100->G2	A100->MI250	
granite_20b			
llama3_8b		AIU2->MI250	A100->G2
granite_34b	2xH100->MI250	2xH100->MI250	
mixtral_8_7b	2xG2->2xMI250	2xMI300X->G2	
llama_70b			

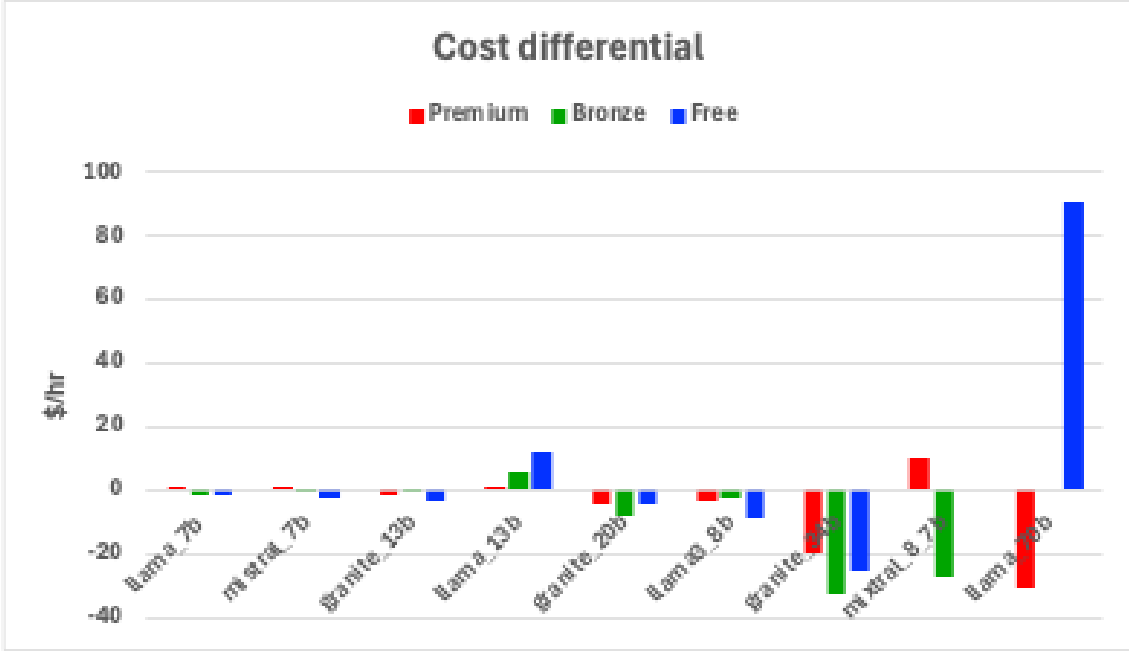


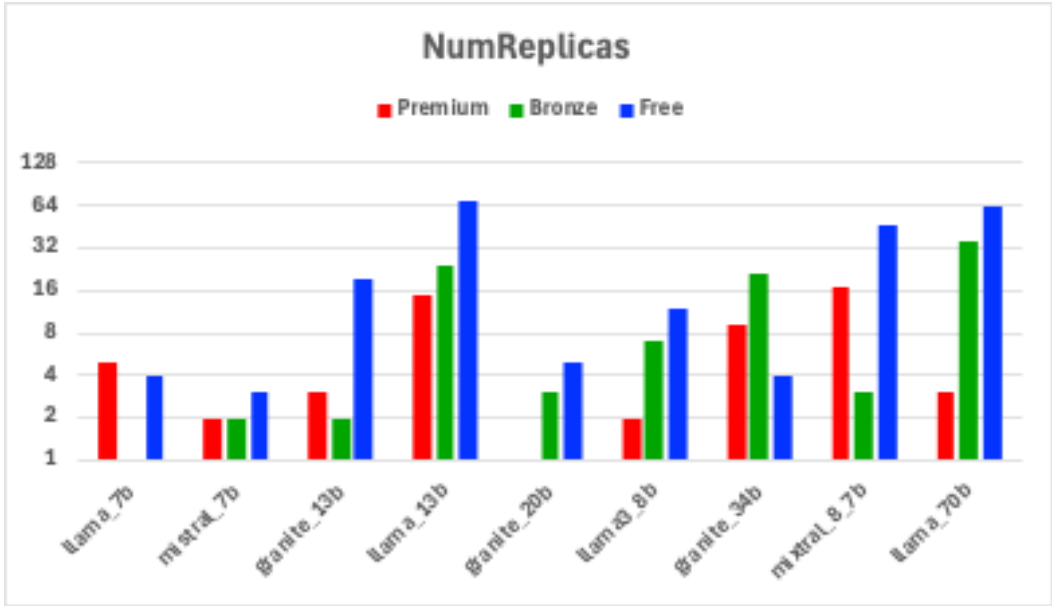
TotalCost
37,272.00

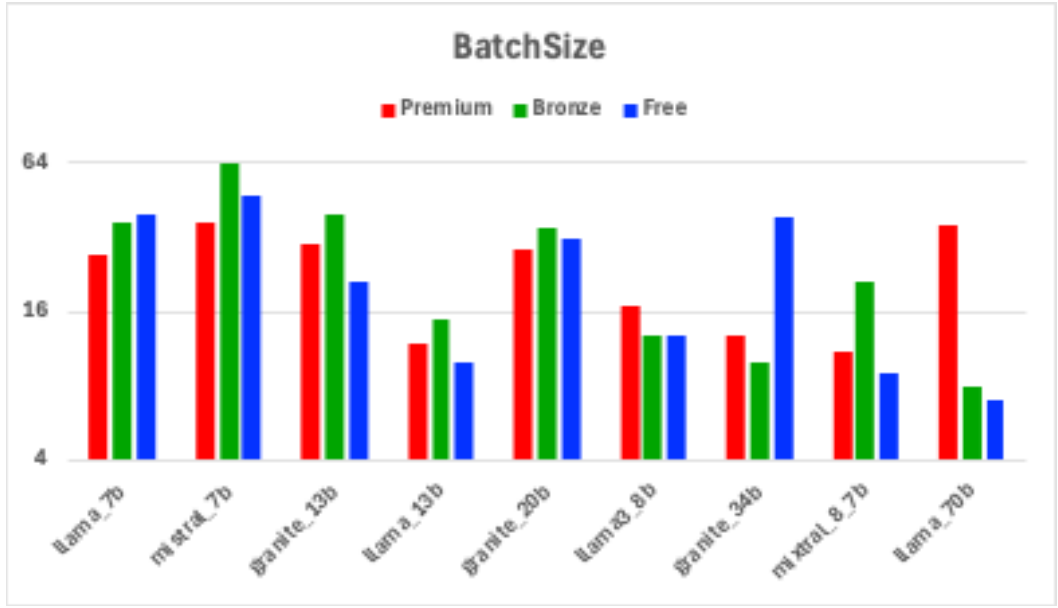
Accelerator Change			
	Premium	Bronze	Free
llama_7b	MI210->MI250	MI210->G2	MI210->MI250
mistral_7b		MI210->MI250	
granite_13b		MI300X->MI250	A100->G2
llama_13b	A100->G2	A100->MI250	
granite_20b			
llama3_8b		AIU2->MI250	A100->G2
granite_34b	2xH100->MI250	2xH100->MI250	
mixtral_8_7b	2xG2->2xMI250	2xMI300X->G2	
llama_70b			

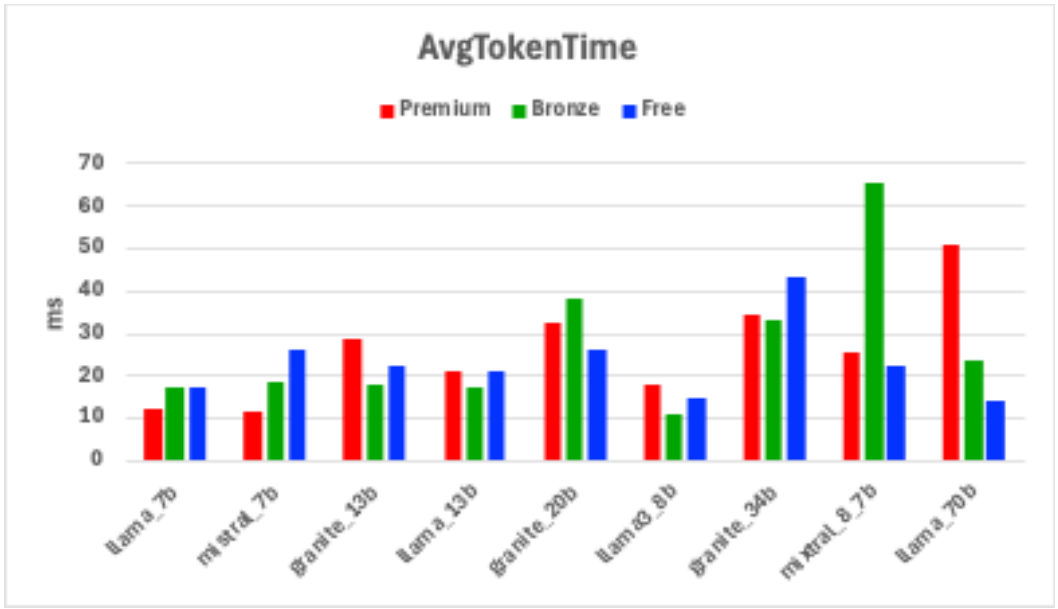
Scale			
	Premium	Bronze	Free
llama_7b			
mistral_7b	1		-9
granite_13b	-3		
llama_13b			51
granite_20b	-11	-19	-10
llama3_8b	-8		
granite_34b			-28
mixtral_8_7b			0
llama_70b	-17	0	35

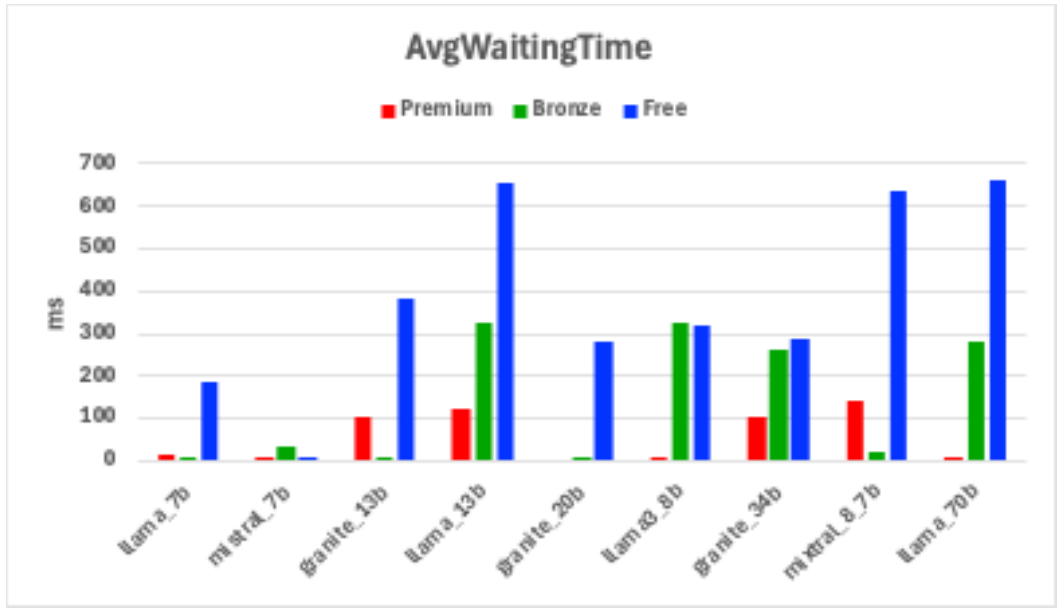





NumReplicas
Premium
Bronze
Free
128
64
32
16
8
4
2
1
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b


BatchSize
Premium
Bronze
Free
64
16
4
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b


AvgTokenTime
Premium
Bronze
Free
ms
70
60
50
40
30
20
10
0
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b

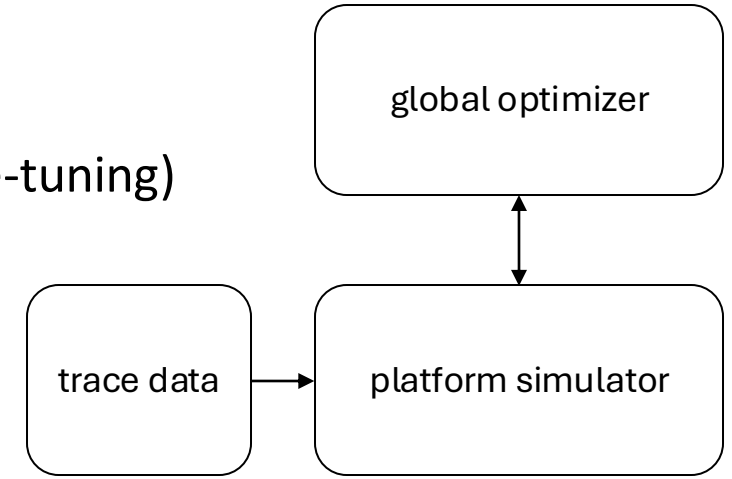

AvgWaitingTime
Premium
Bronze
Free
ms
700
600
500
400
300
200
100
0
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b

Global Optimization

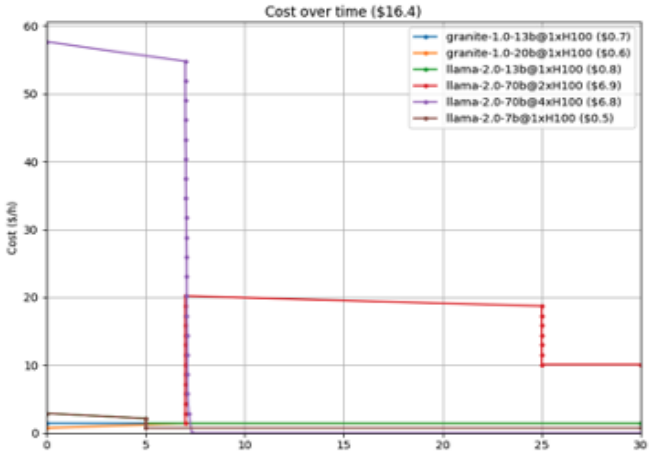
Initial Results/Gains

Global optimization results on simulated workload

1. Satisfy same performance (ITL) with less GPUs (91xH100 -> 37xH100)
2. Spare GPUs could be utilized by other workloads (batch inference/fine-tuning)
3. 2x financial gain



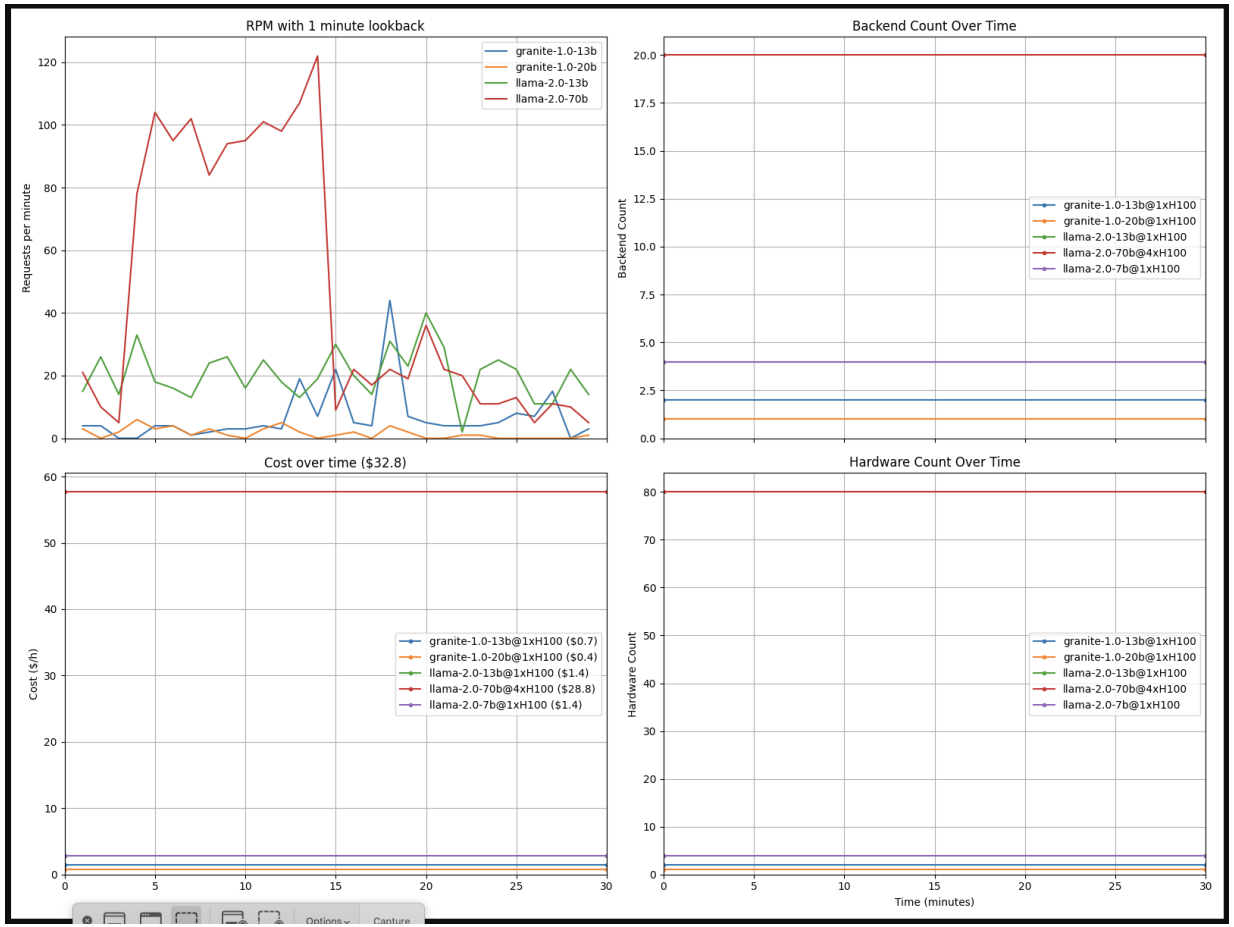
***No** optimization*
($ 32.8)



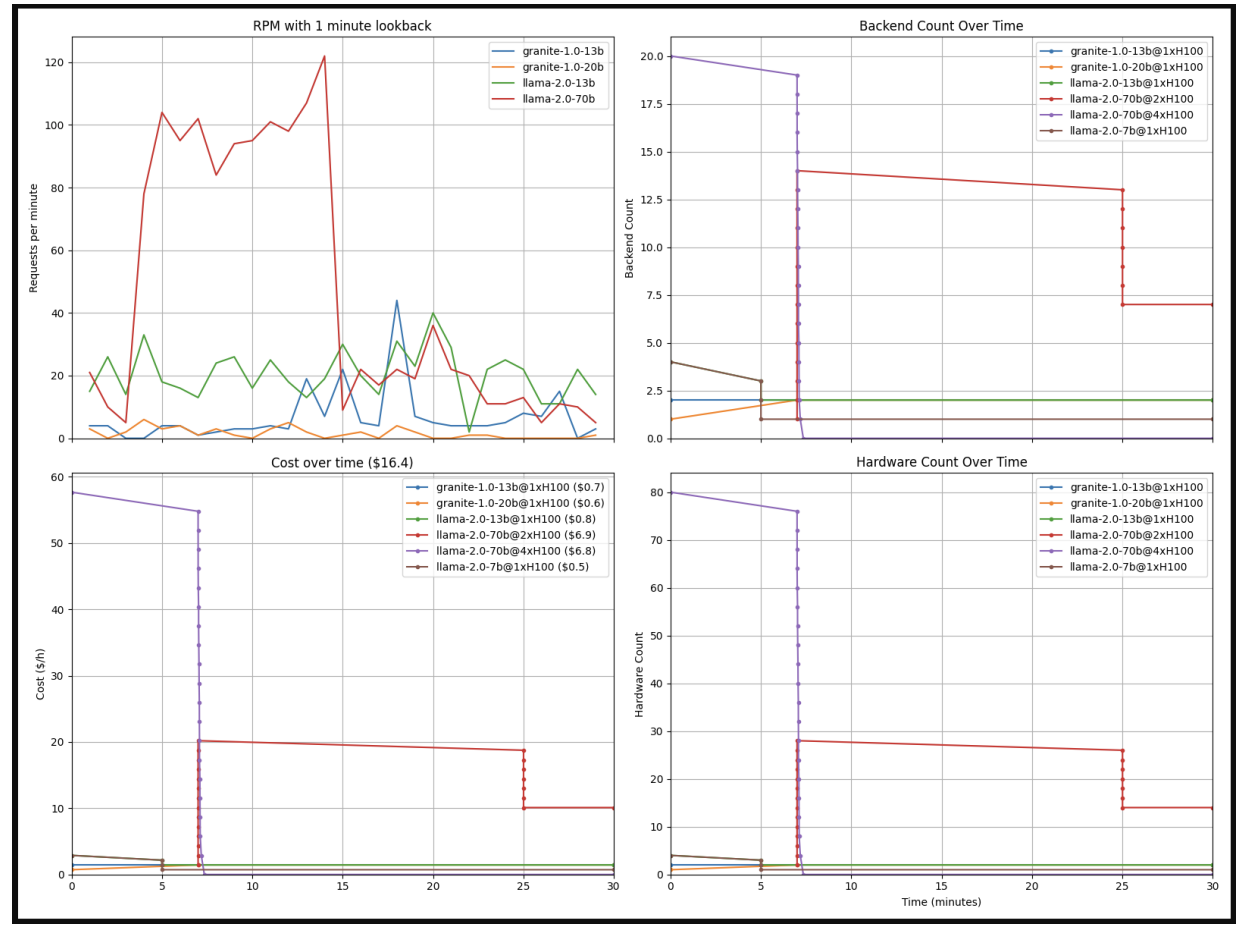
***With** optimization*
($ 16.4)

model	GPU	replicas
granite-1.0-13b	1xH100	2 -> 2
granite-1.0-20b	1xH100	1 -> 2
llama-2.0-13b	1xH100	4 -> 2
llama-2.0-70b	4xH100 -> 2xH100	20 -> 14
llama-2.0-7b	1xH100	4 -> 3

optimizer off



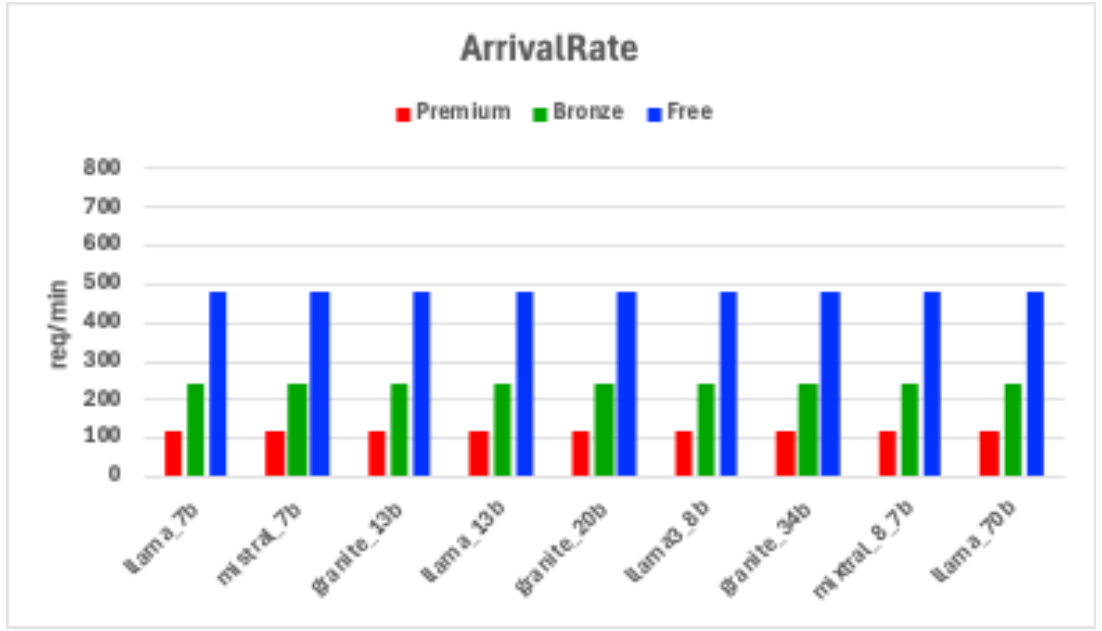
optimizer on
5 min - tick
2 min - scale up

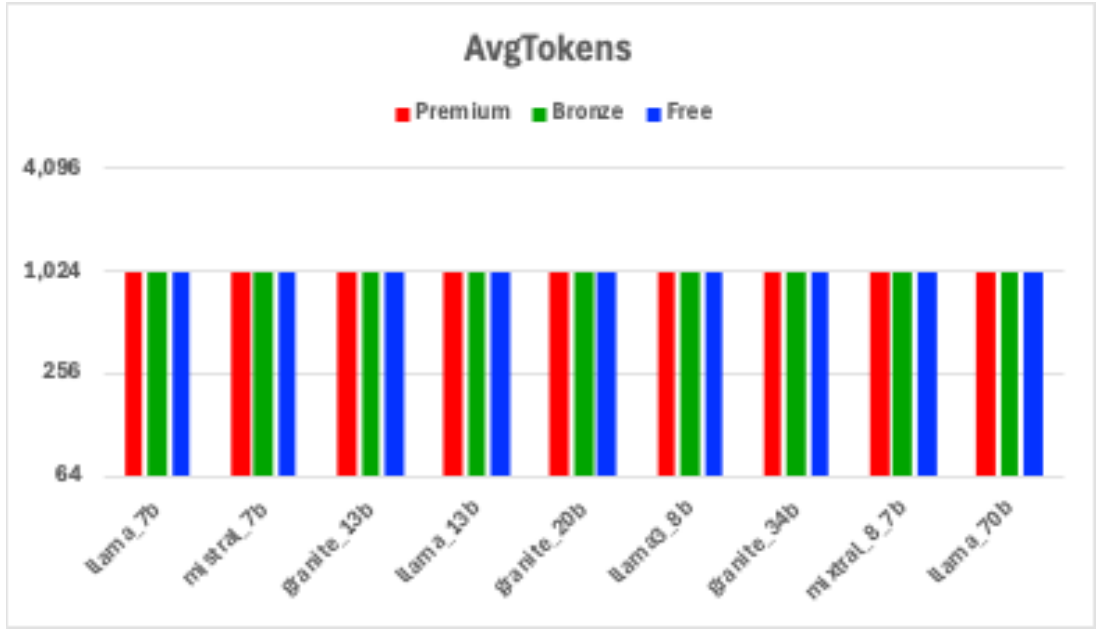


Backup

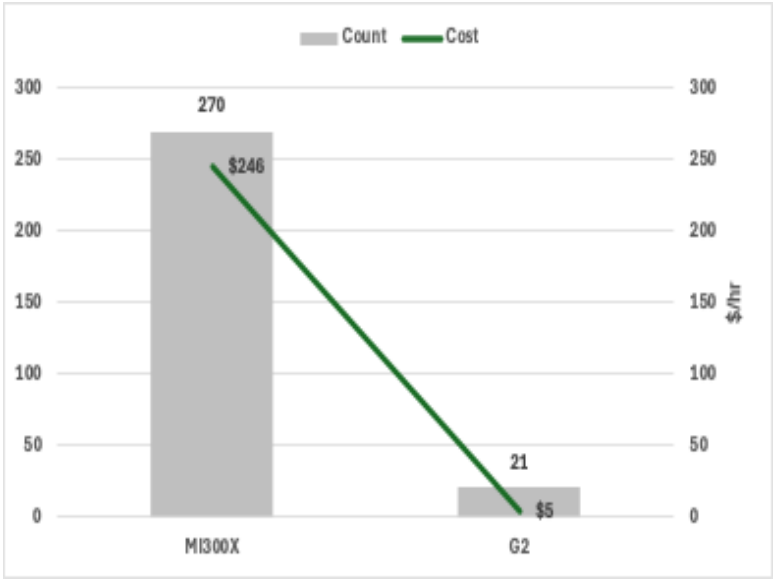
Unlimited accelerators

- capacity planning
- cloud deployment
- separable optimization

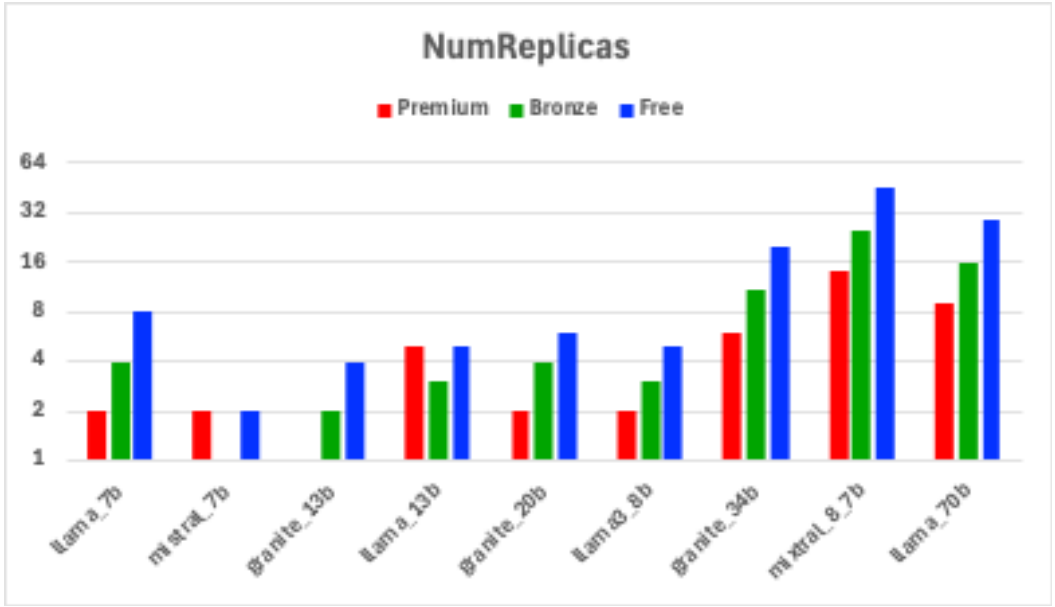


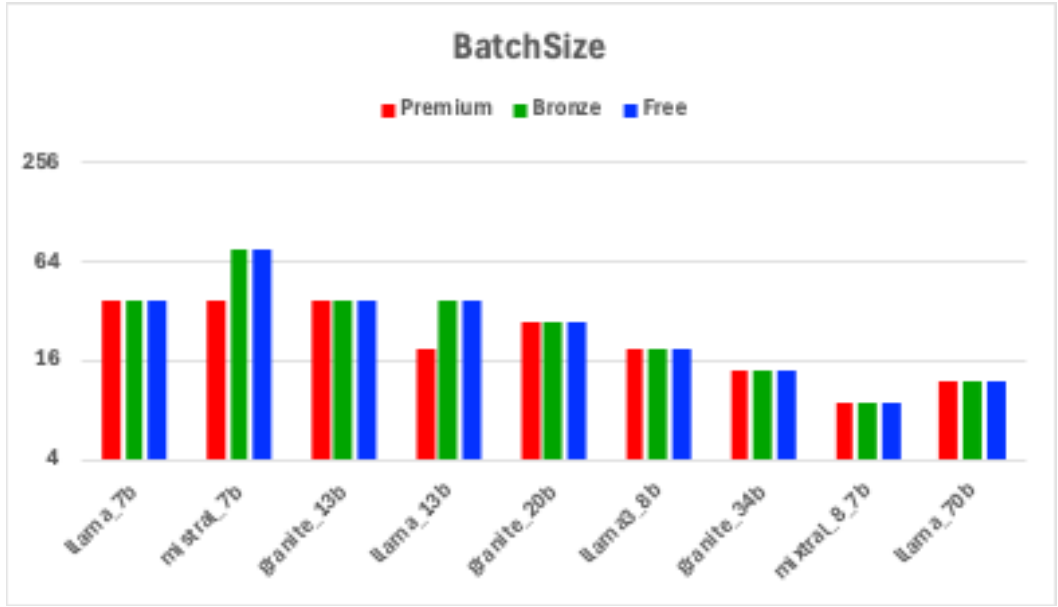


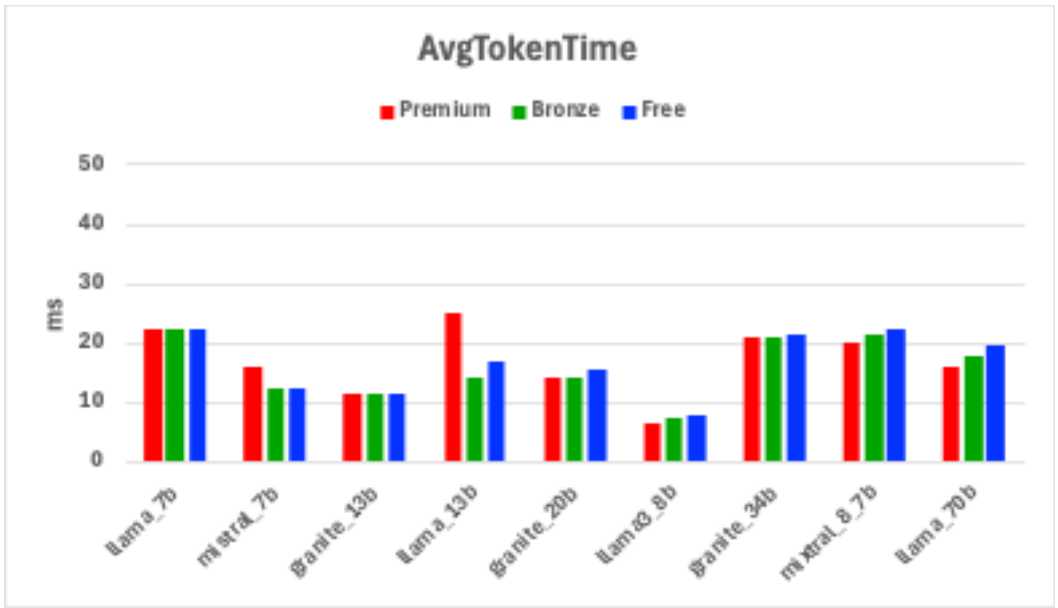
Accelerator			
	Premium	Bronze	Free
llama_7b	G2	G2	G2
mistral_7b	G2	MI300X	MI300X
granite_13b	MI300X	MI300X	MI300X
llama_13b	G2	MI300X	MI300X
granite_20b	MI300X	MI300X	MI300X
llama3_8b	MI300X	MI300X	MI300X
granite_34b	MI300X	MI300X	MI300X
mixtral_8_7b	MI300X	MI300X	MI300X
llama_70b	2xMI300X	2xMI300X	2xMI300X

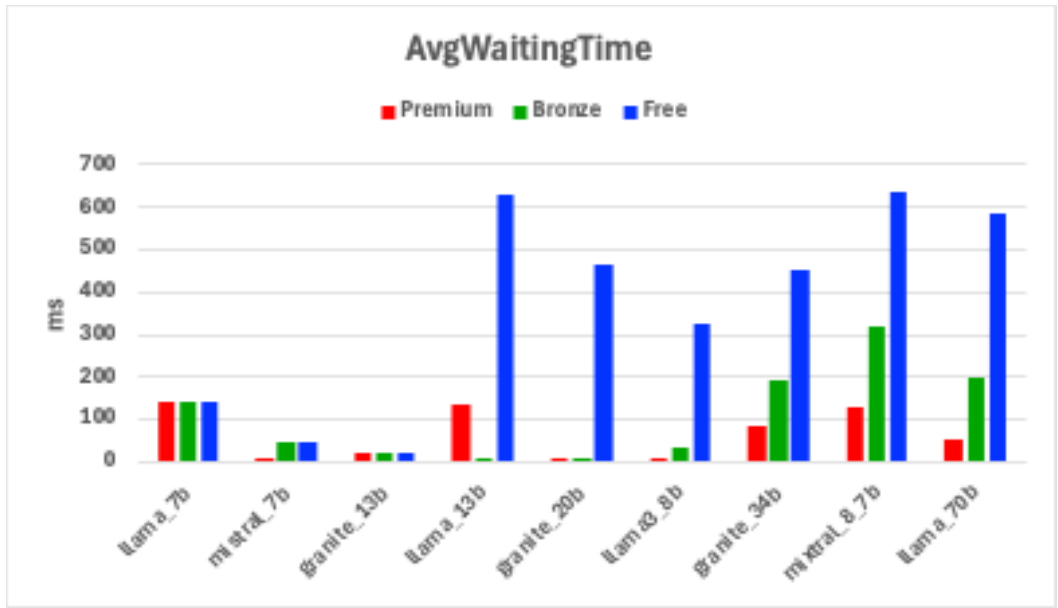


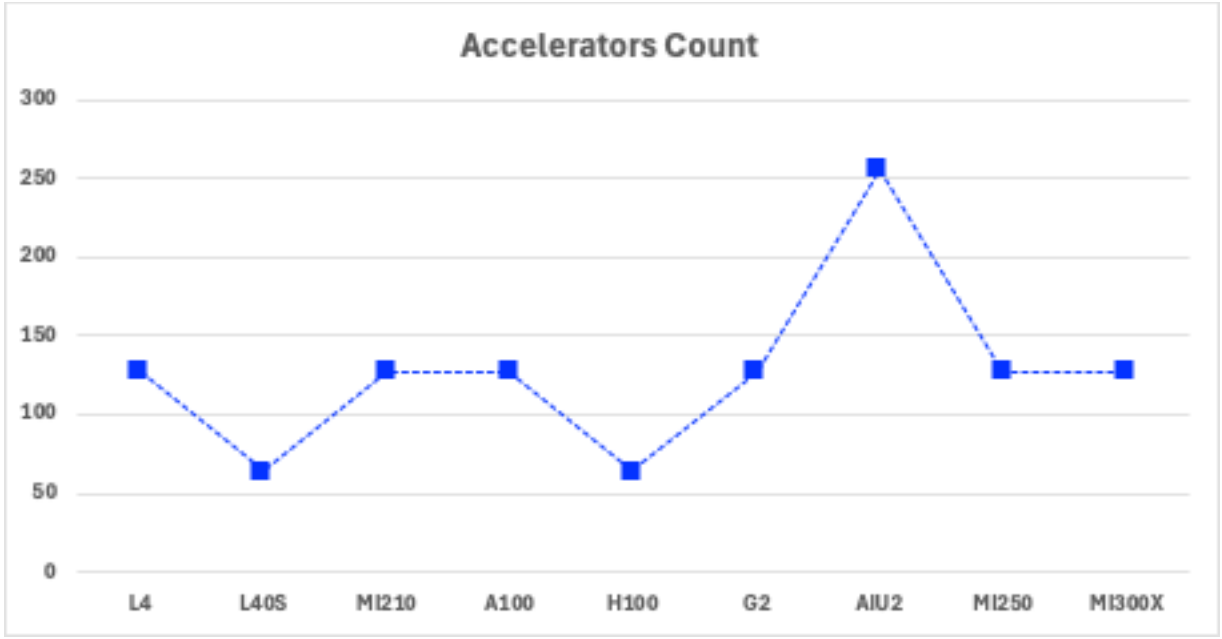
TotalCost
25,053.00


NumReplicas
Premium
Bronze
Free
64
32
16
8
4
2
1
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b


BatchSize
Premium
Bronze
Free
256
64
16
4
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b

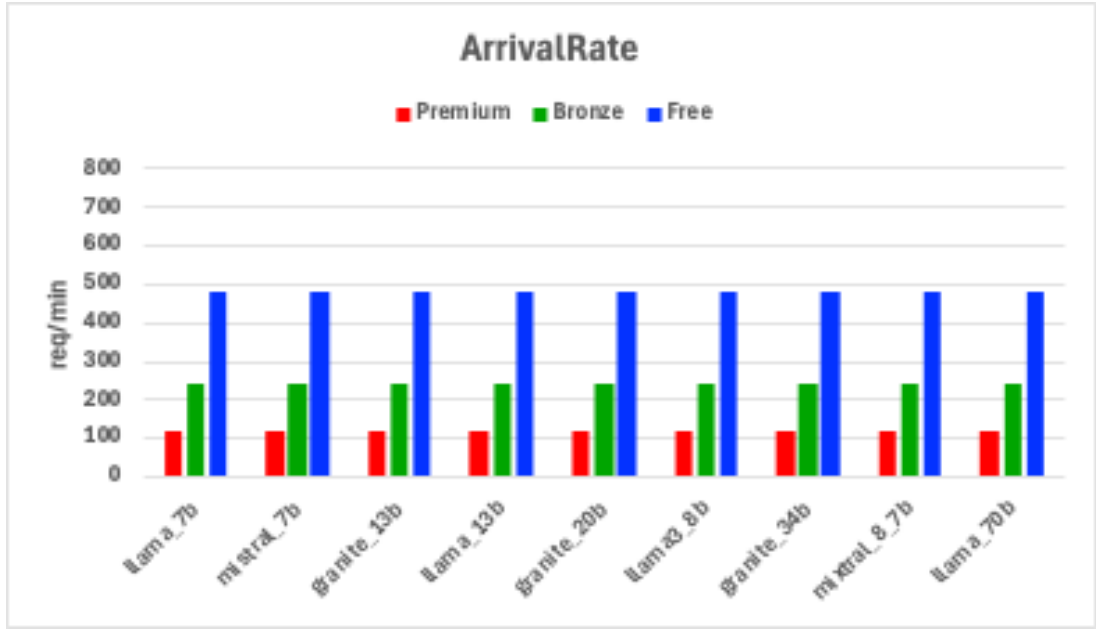

AvgTokenTime
Premium
Bronze
Free
ms
50
40
30
20
10
0
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b

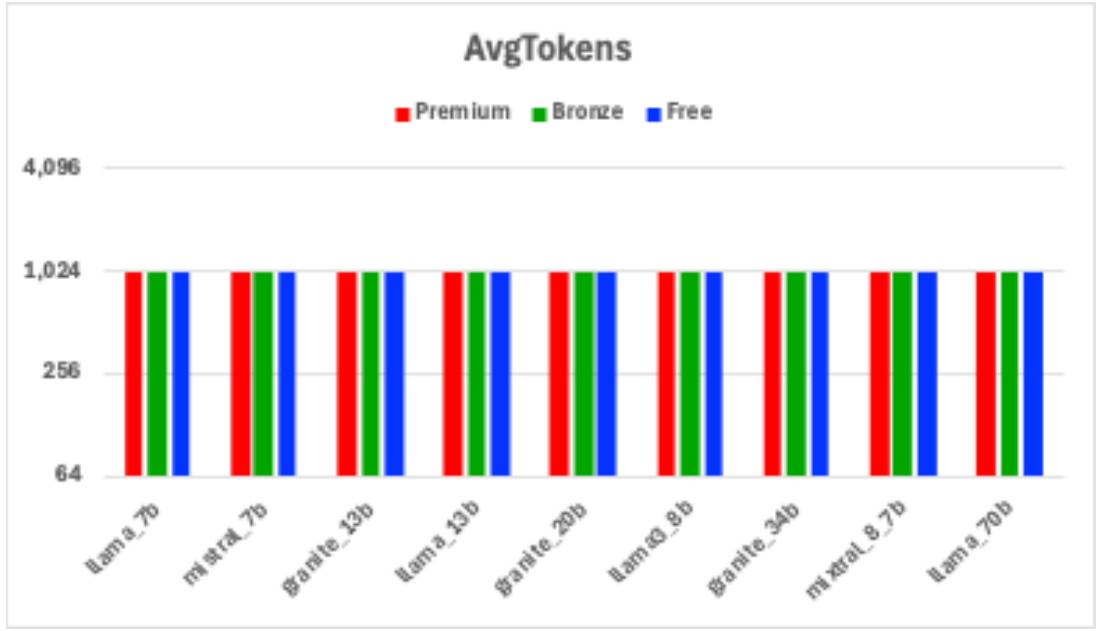

AvgWaitingTime
Premium
Bronze
Free
ms
700
600
500
400
300
200
100
0
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b



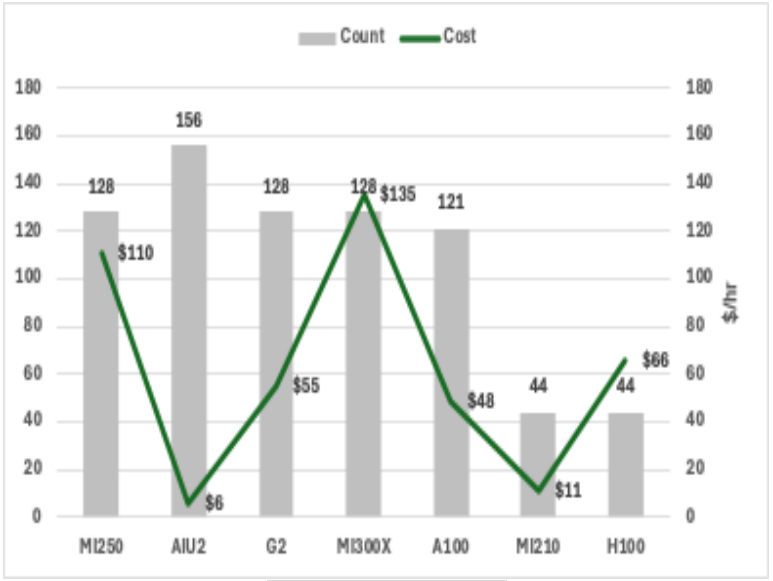
Limited accelerators

- cluster deployment
- greedy optimization

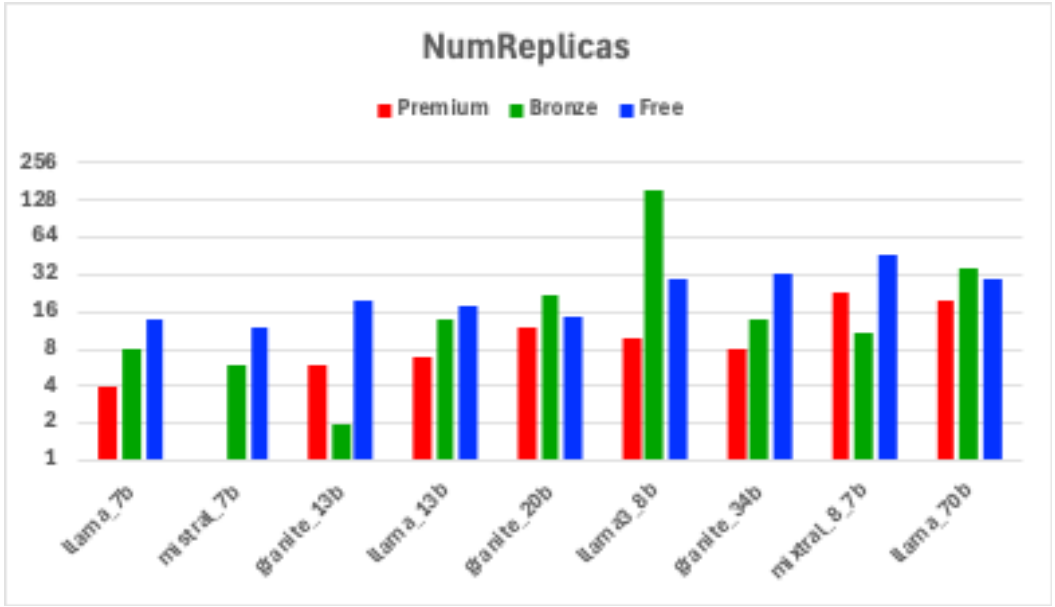


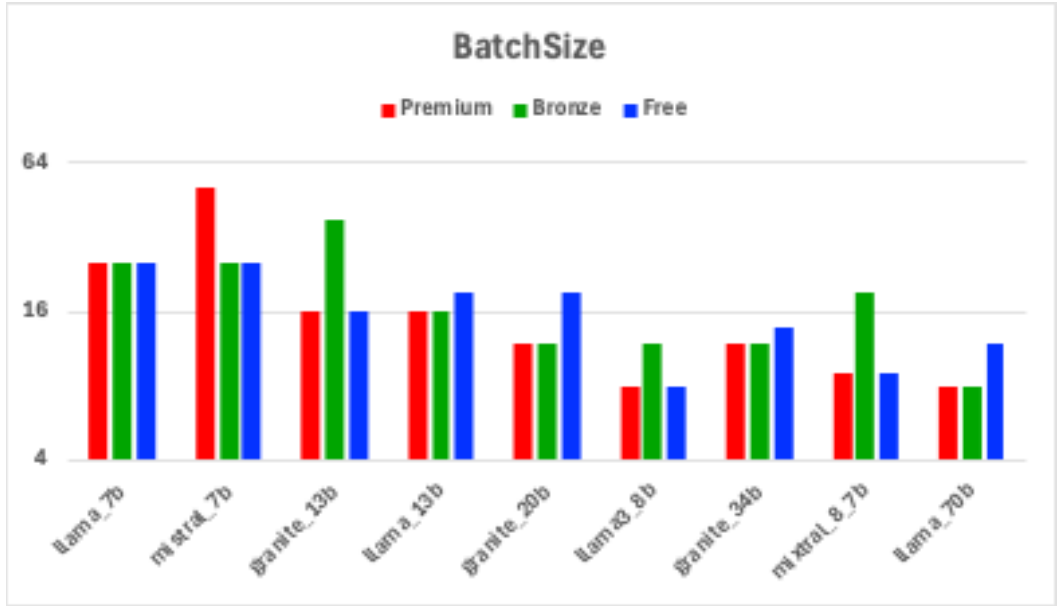


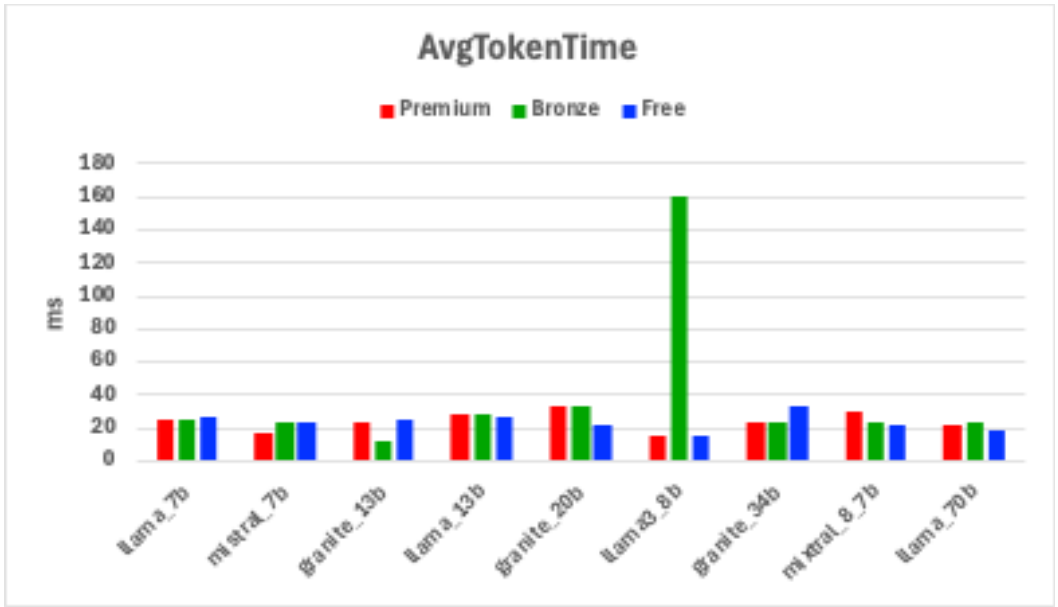
Accelerator			
	Premium	**Bronze**	**Free**
llama_7b	MI210	MI210	MI210
mistral_7b	MI250	MI210	MI210
granite_13b	A100	MI300X	A100
llama_13b	A100	A100	G2
granite_20b	A100	A100	MI250
llama3_8b	A100	AIU2	A100
granite_34b	2xH100	2xH100	2xG2
mixtral_8_7b	2xG2	2xMI300X	MI300X
llama_70b	2xMI250	2xMI250	2xMI300X

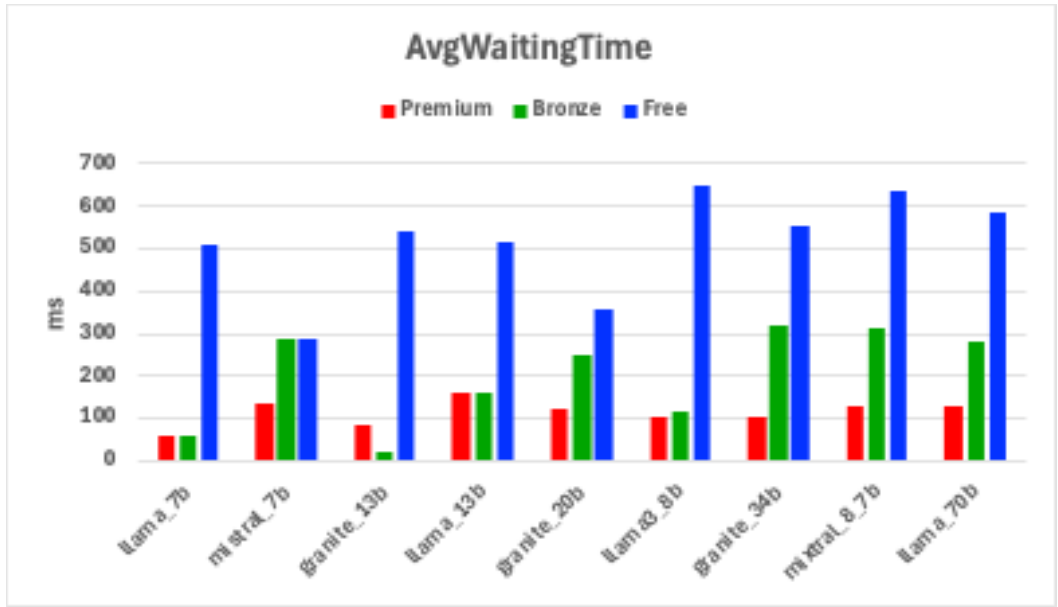


TotalCost
43,203.00


NumReplicas
Premium
Bronze
Free
256
128
64
32
16
8
4
2
1
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b


BatchSize
Premium
Bronze
Free
64
16
4
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b


AvgTokenTime
Premium
Bronze
Free
ms
180
160
140
120
100
80
60
40
20
0
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b


AvgWaitingTime
Premium
Bronze
Free
ms
700
600
500
400
300
200
100
0
llama_7b
mistral_7b
granite_13b
llama_13b
granite_20b
llama3_8b
granite_34b
mixtral_8_7b
llama_70b