Discussion on Feb21

$\boxed{step1}$ running `vllm`

$\boxed{step2}$ `flash attention` kernel api entry point

$\boxed{step3}$ `page` / `no cache` input runtime

Process by Mar3

$\boxed{step1}$ running `vllm`

Done

$\boxed{step2}$ `flash attention` kernel api entry point

`Method1` from the `vllm` repo. 【WIP】

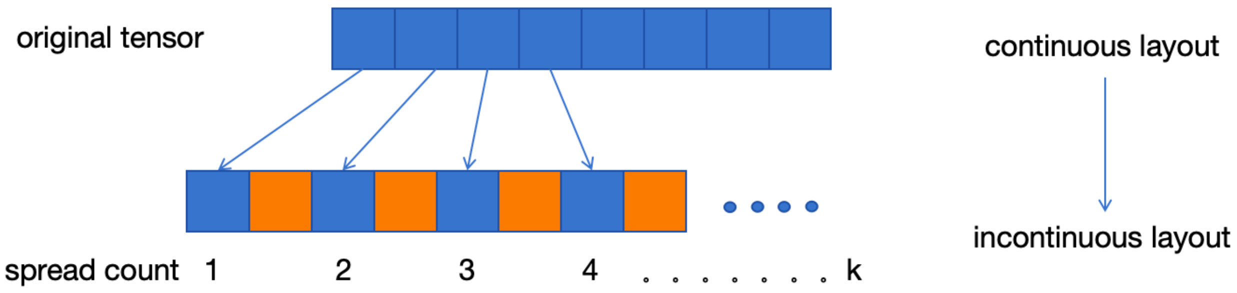
`vllm/attention/layer.py`and`vllm/attention/selector.py`

`Method2`from the `flash-attention`repo

More easy but unknown about efficiency

```
from flash_attn import flash_attn_func, flash_attn_qkvpacked_func, flash_attn_with_kvcache
```

$\boxed{step3}$ `page` / `no cache` input runtime



```
prompts = [
    "Hello, my name is",
    "The president of the United States is",
    "The capital of France is",
    "The future of AI is",
]
```

`Page cache runtime`:

```
Task 1 runtime: 0.0753 seconds
Prompt: 'Hello, my name is'
Generated text: ' Joel, my dad is my friend and we are in
a relationship. I am'

Prompt: 'The president of the United States is'
Generated text: ' speaking out against the release of some
State Department documents which show the Russians were
involved'

Prompt: 'The capital of France is'
Generated text: ' known as the “Proud French capital”.
What is this city'

Prompt: 'The future of AI is'
Generated text: ' literally in danger of being taken by
any other company.\nAgreed. '
```

`No cache runtime`:

```
for prompt in prompts:
    llm_task2.reset_prefix_cache() #delete cache manually
```

```
for prompt in prompts:
    a = time.time()
    llm_task2.reset_prefix_cache()
    b = time.time()
    print(f"clear runtime: {b - a:.4f} seconds")
```

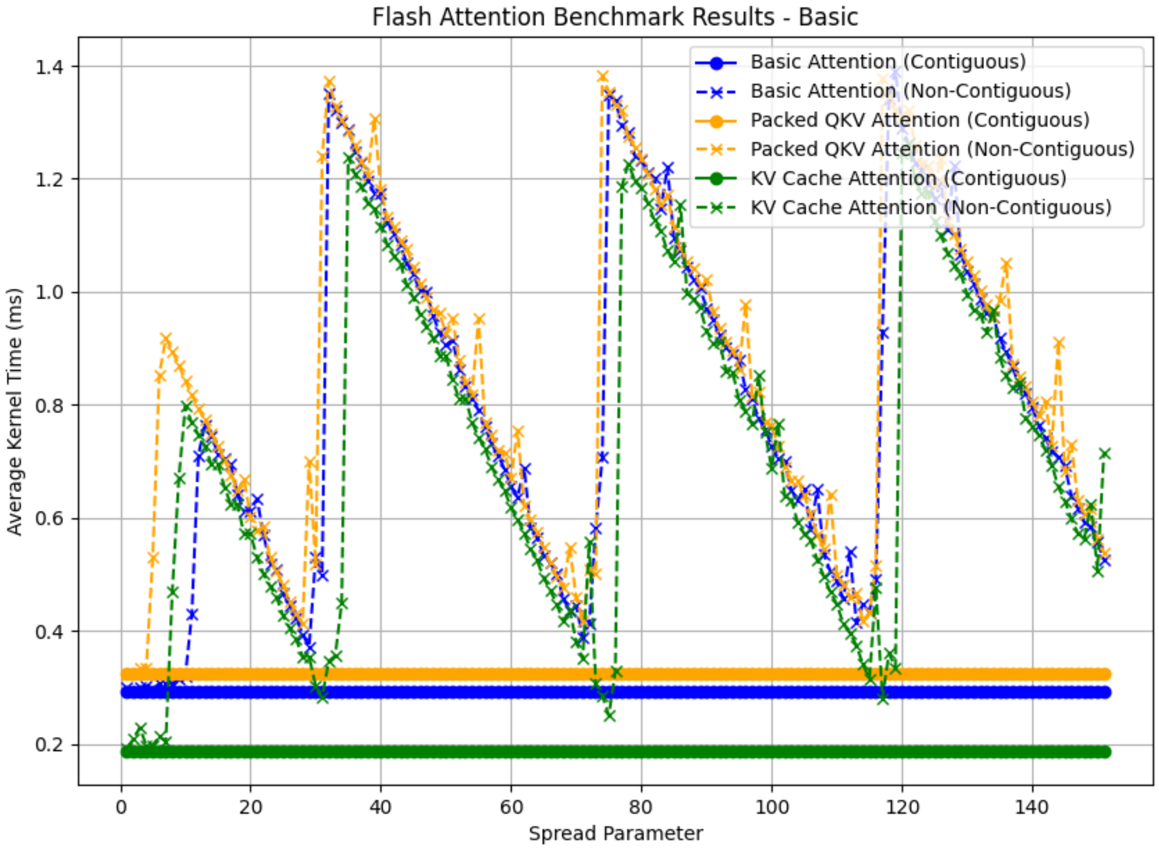
```
clear runtime: 0.0000 seconds
```

```
Task 2 runtime: 0.1552 seconds
Prompt: 'Hello, my name is'
Generated text: ' Joel, my dad is my friend and we are in
a relationship. I am'

Prompt: 'The president of the United States is'
Generated text: ' using a woman as a political pawn for
the US, and it’s'

Prompt: 'The capital of France is'
Generated text: " a far cry from this one.  It's a city
with countless cultures and"

Prompt: 'The future of AI is'
Generated text: ' we are going to have the most
intelligent human being on earth.   It'
```



```
Comparison Results (spread=1180):
----------------------------------------------------------
----------------------
Benchmark / Layout        |    Warm-up Time (ms) |   Core
Avg Time (ms)
----------------------------------------------------------
----------------------
Basic Attention
               Contiguous |              595.341 |
0.122
           Non-contiguous |                0.462 |
0.134
                Core Diff |                      |
+0.011 ( +9.06%)
```

```
----------------------------------------------------------
----------------------
Packed QKV Attention
               Contiguous |                0.291 |
0.149
           Non-contiguous |                0.198 |
0.158
                Core Diff |                      |
+0.009 ( +5.73%)
----------------------------------------------------------
----------------------
KV Cache Attention
               Contiguous |                1.490 |
0.034
           Non-contiguous |                0.051 |
0.044
                Core Diff |                      |
+0.010 (+30.82%)
----------------------------------------------------------
----------------------
```