# Some Optimization Strategies in Milvus

Heterogeneous Computing

董若扬

2024-10-02

1. CPU-oriented Optimizations	2
1.1 Background	3
1.2 How Milvus Addresses These?	1
1.3 Cache-aware Design in Milvus5	5
2. GPU-oriented Optimizations	5
2.1 Supporting bigger k in GPU kernel	7
2.2 Supporting multi-GPU devices	3
3. GPU AND CPU CO-DESIGN	)
3.1 The Limitations	)
3.2 Addressing the first limitation	1
3.3 Addressing the second limitation	2

1. CPU-oriented Optimizations	2
1.1 Background	. 3
1.2 How Milvus Addresses These?	. 4
1.3 Cache-aware Design in Milvus	. 5
2. GPU-oriented Optimizations	
2.1 Supporting bigger k in GPU kernel	. 7
2.2 Supporting multi-GPU devices	. 8
3. GPU and CPU Co-design	
3.1 The Limitations	
3.2 Addressing the first limitation.	11
3.3 Addressing the second limitation.	12

CPU-ORIENTED OPTIMIZATIONS

GPU-ORIENTED OPTIMIZATIONS

GPU AND CPU Co-DESIGN

 $\circ$ 

# 1.1 Background

#### 1.2 How Milvus Addresses These?

 

#### 1.3 Cache-aware Design in Milvus

1. CPU-oriented Optimizations	2
1.1 Background	3
1.2 How Milvus Addresses These?	
1.3 Cache-aware Design in Milvus	5
2. GPU-oriented Optimizations	6
2.1 Supporting bigger k in GPU kernel	7
2.2 Supporting multi-GPU devices	8
3. GPU and CPU Co-design	
3.1 The Limitations	10
3.2 Addressing the first limitation.	11
3.3 Addressing the second limitation	12

GPU AND CPU Co-design

○
○

#### 2.1 Supporting bigger k in GPU kernel

GPU AND CPU Co-DESIGN

0

## 2.2 Supporting multi-GPU devices

1. CPU-oriented Optimizations	. 2
1.1 Background	. 3
1.2 How Milvus Addresses These?	. 4
1.3 Cache-aware Design in Milvus	. 5
2. GPU-oriented Optimizations	. 6
2.1 Supporting bigger k in GPU kernel	
2.2 Supporting multi-GPU devices	. 8
3. GPU AND CPU CO-DESIGN	9
3.1 The Limitations	10
3.2 Addressing the first limitation	1
3.3 Addressing the second limitation	12

CPU-oriented Optimizations

GPU-oriented Optimizations

GPU AND CPU Co-design

•

0

3.1 The Limitations

GPU-oriented Optimizations

GPU AND CPU Co-DESIGN

•

3.2 Addressing the first limitation.

0

### 3.3 Addressing the second limitation.