Some Optimization Strategies in Milvus

Heterogeneous Computing

董若扬

2024-10-02

ADSLAB, USTC

Outline

Outline

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 7

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. 11

3.3 Addressing the second limitation. 12

1.1 Background

1.2 How Milvus Addresses These?

1.3 Cache-aware Design in Milvus

Outline

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 7

    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. 11

    3.3 Addressing the second limitation. 12

2.1 Supporting bigger k in GPU kernel

2.2 Supporting multi-GPU devices

Outline

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 7

    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. 11

    3.3 Addressing the second limitation. 12

3.1 The Limitations

3.2 Addressing the first limitation.

3.3 Addressing the second limitation.