Windows Kernel Internals
User-mode Heap Manager

David B. Probert, Ph.D.

Windows Kernel Development

Microsoft Corporation

Topics

- Common problems with the NT heap
- LFH design
- Benchmarks data
- Heap analysis

Default NT Heap

- Unbounded fragmentation for the worst scenario:
  - External fragmentation
  - Virtual address fragmentation
- Poor performance for:
  - Large heaps
  - SMP
  - Large blocks
  - Fast growing scenarios
  - Fragmented heaps

Goals For LFH

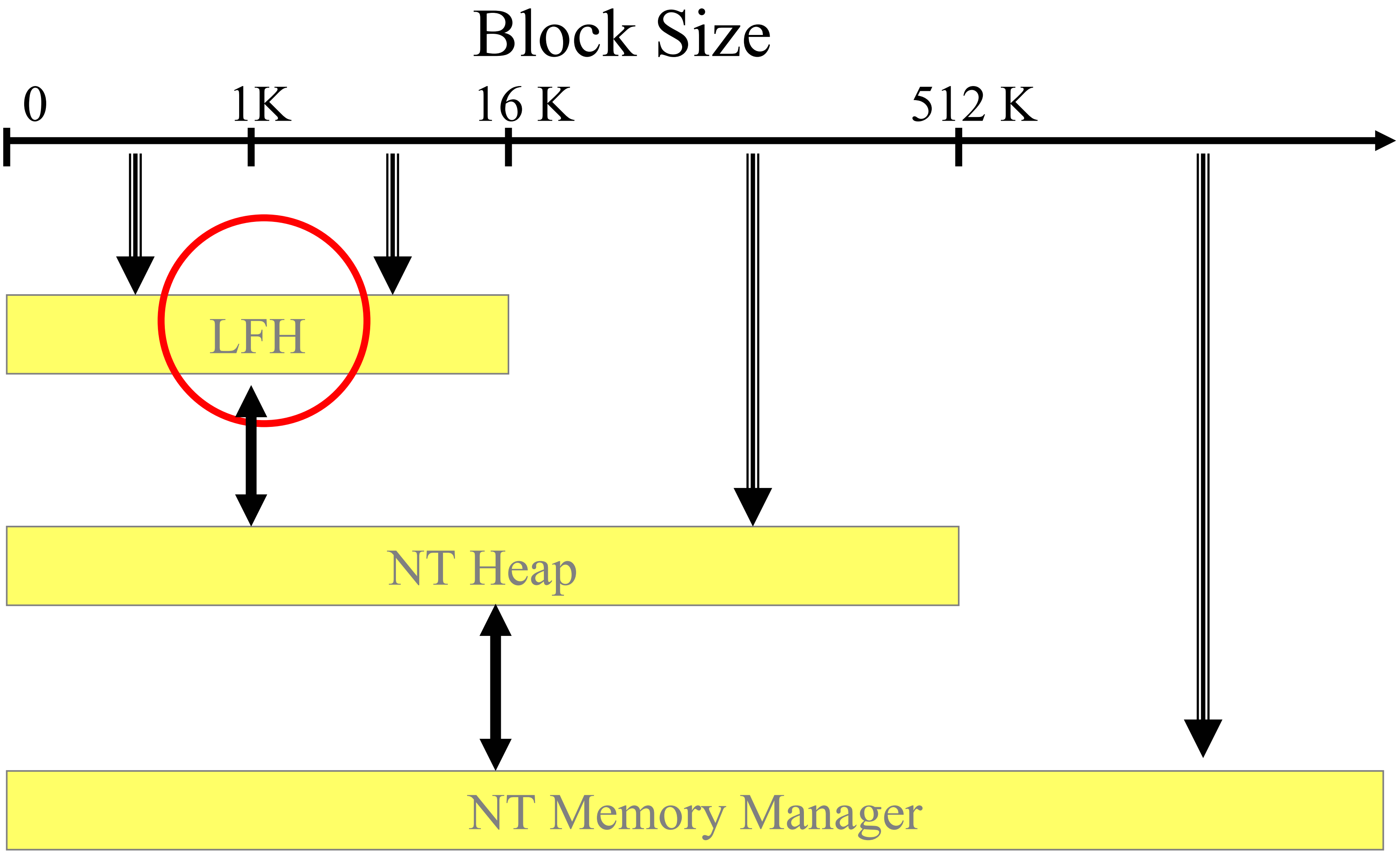
- Bounded low fragmentation
- Low risk (minimal impact)
- Stable and high performance for:
  - Large heaps
  - Large blocks
  - SMP
  - Long running applications

LFH Design

- Bucket-oriented heap
- Better balance between internal and external fragmentation
- Improved data locality
- No locking for most common paths

Tradeoffs

- Performance / footprint
- Internal / external fragmentation
- Thread / processor data locality
- Using prefetch techniques



Block Size

0

1K

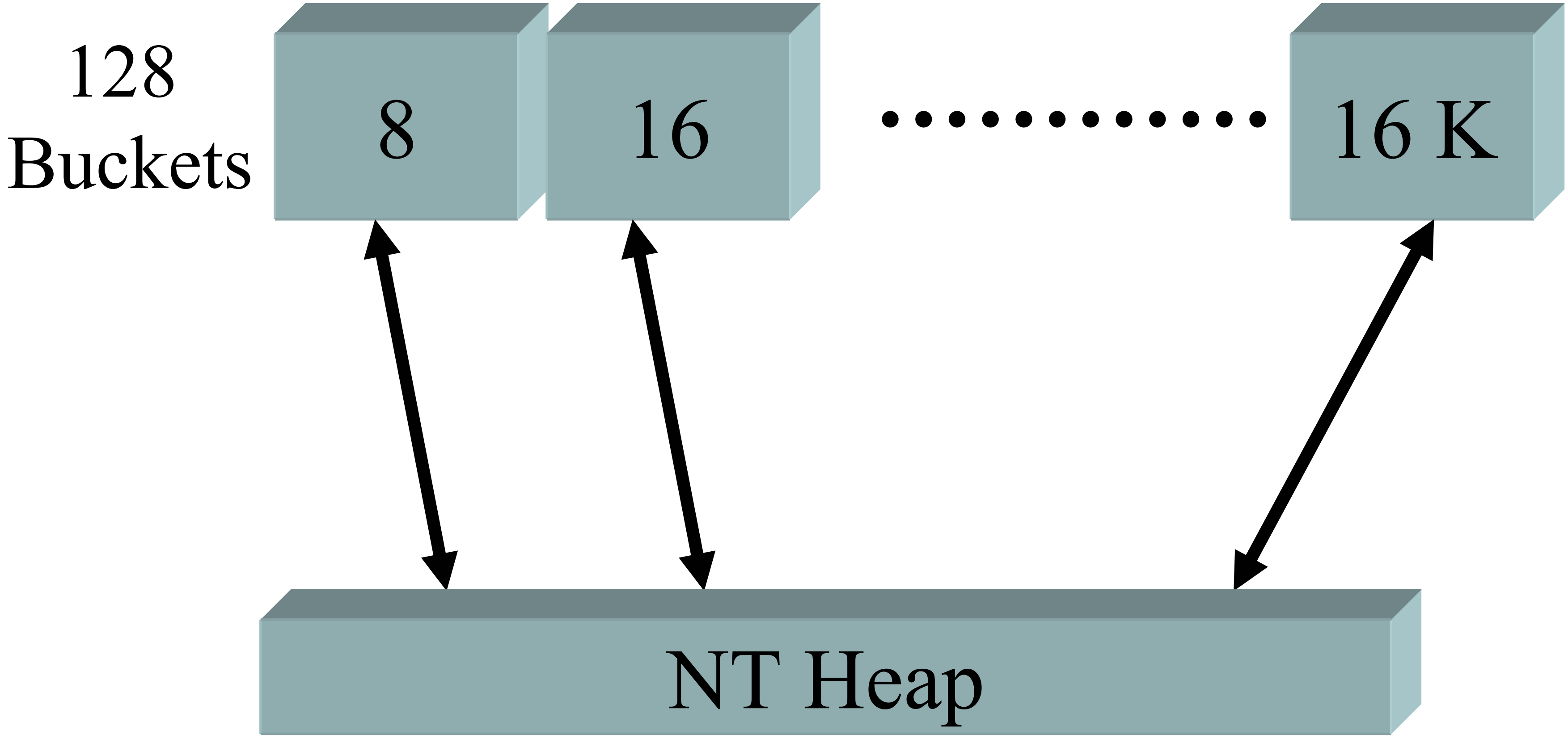
16 K

512 K

LFH

NT Heap

NT Memory Manager


128
Buckets

8

16

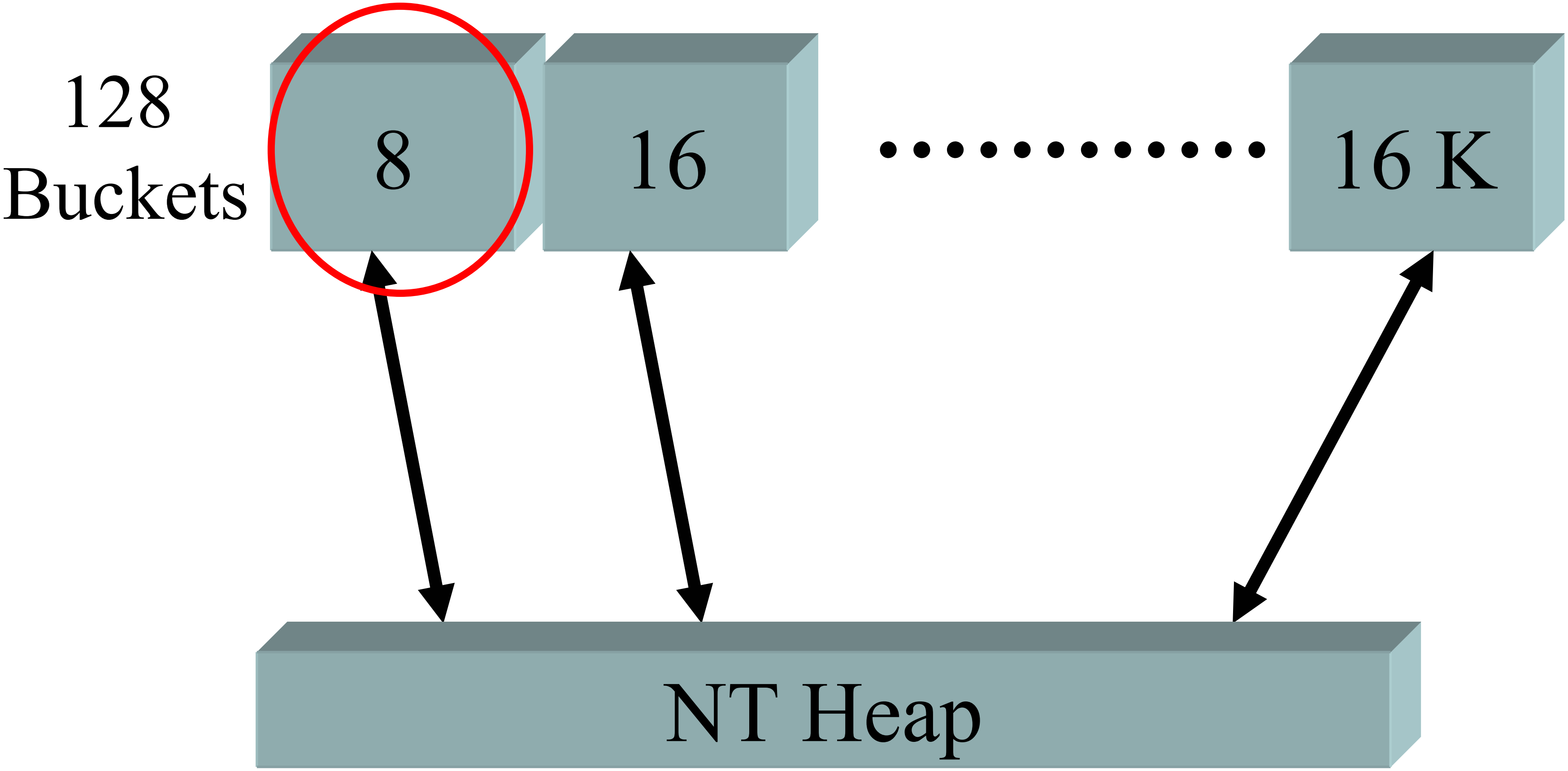
…………

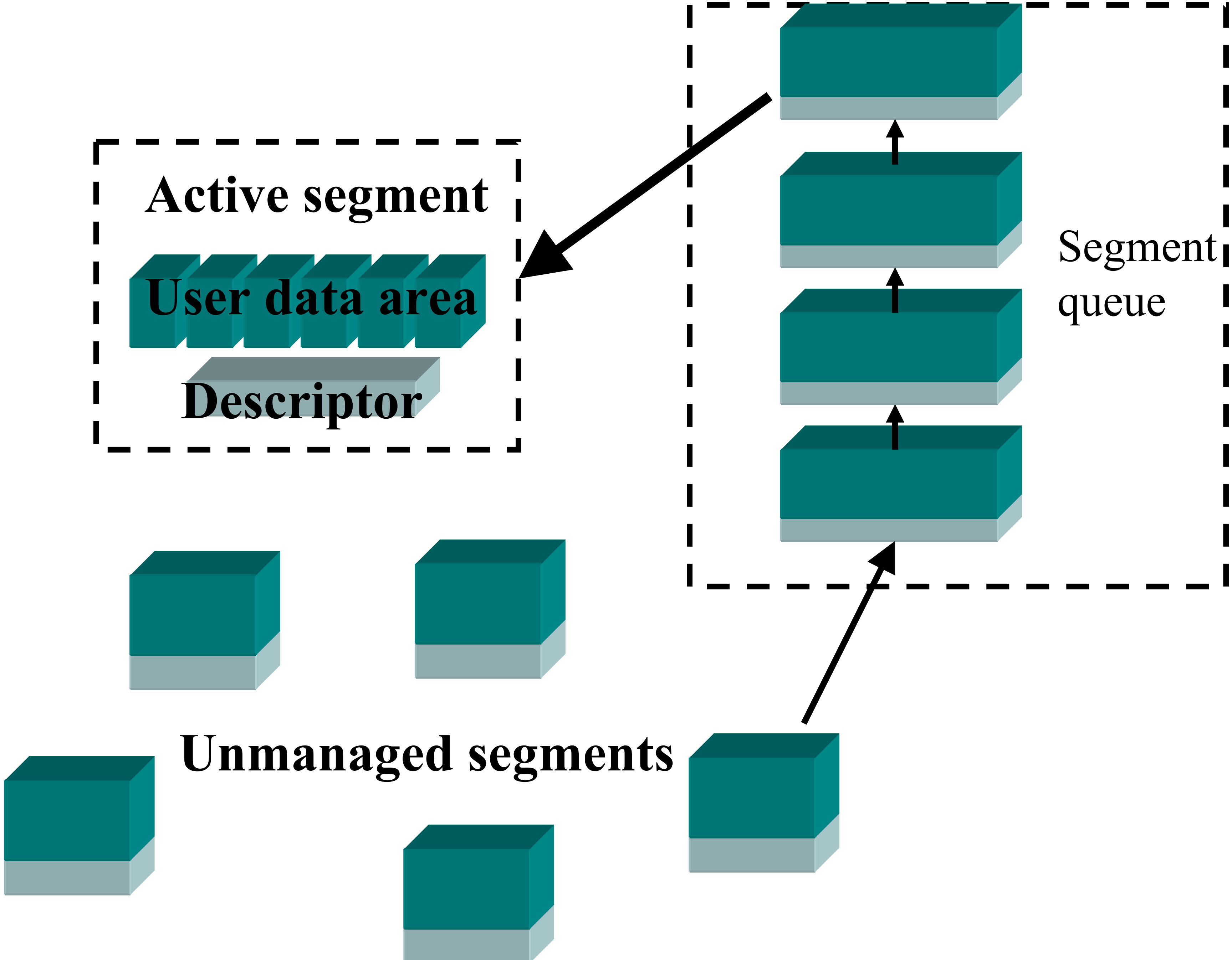
16 K

NT Heap

Allocation Granularity

Block Size	Granularity	Buckets
256	8	32
512	16	16
1024	32	16
2048	64	16
4096	128	16
8196	256	16
16384	512	16


128
Buckets
8
16
16 K
NT Heap



Active segment

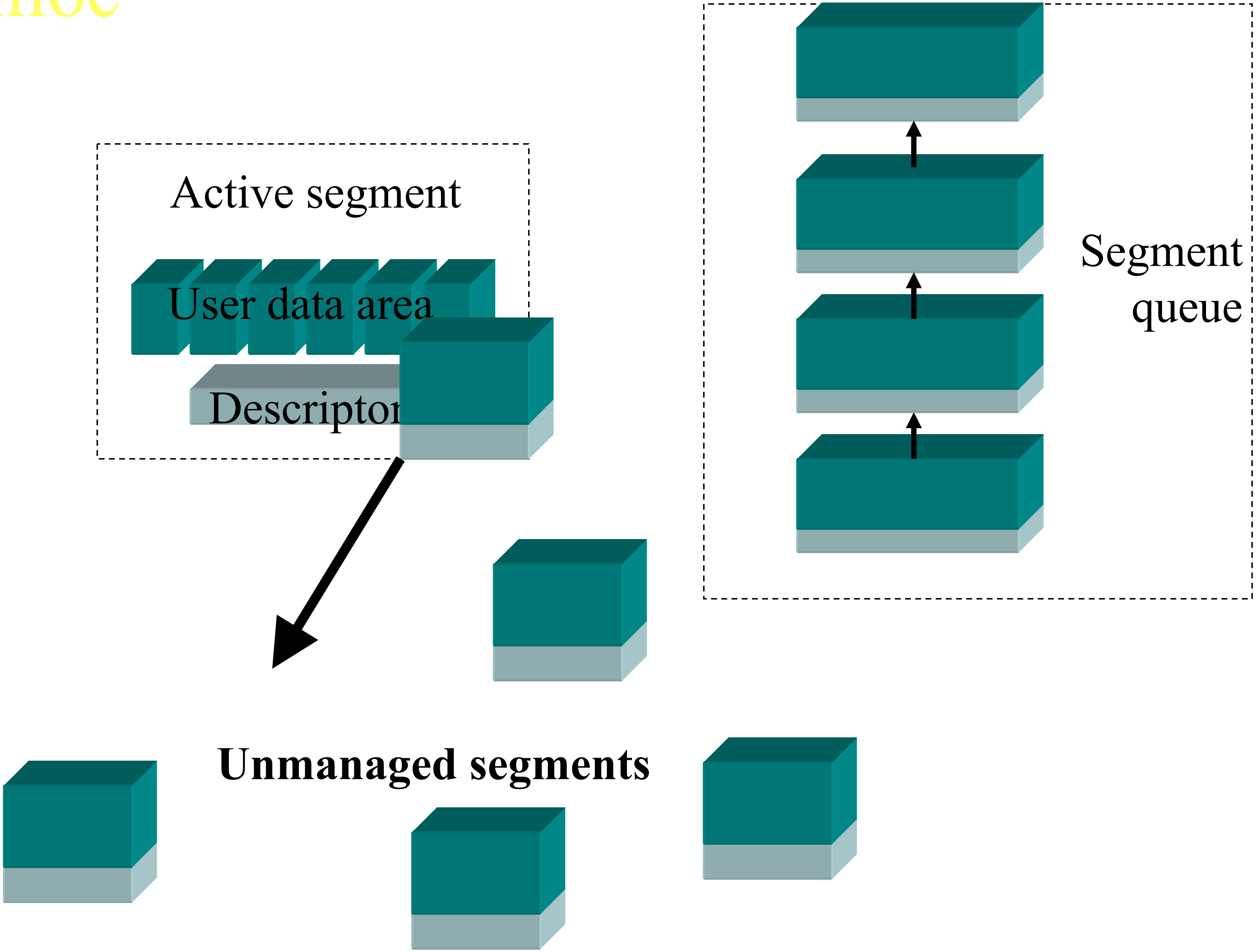
User data area

Descriptor

Segment queue

Unmanaged segments

Alloc



Active segment

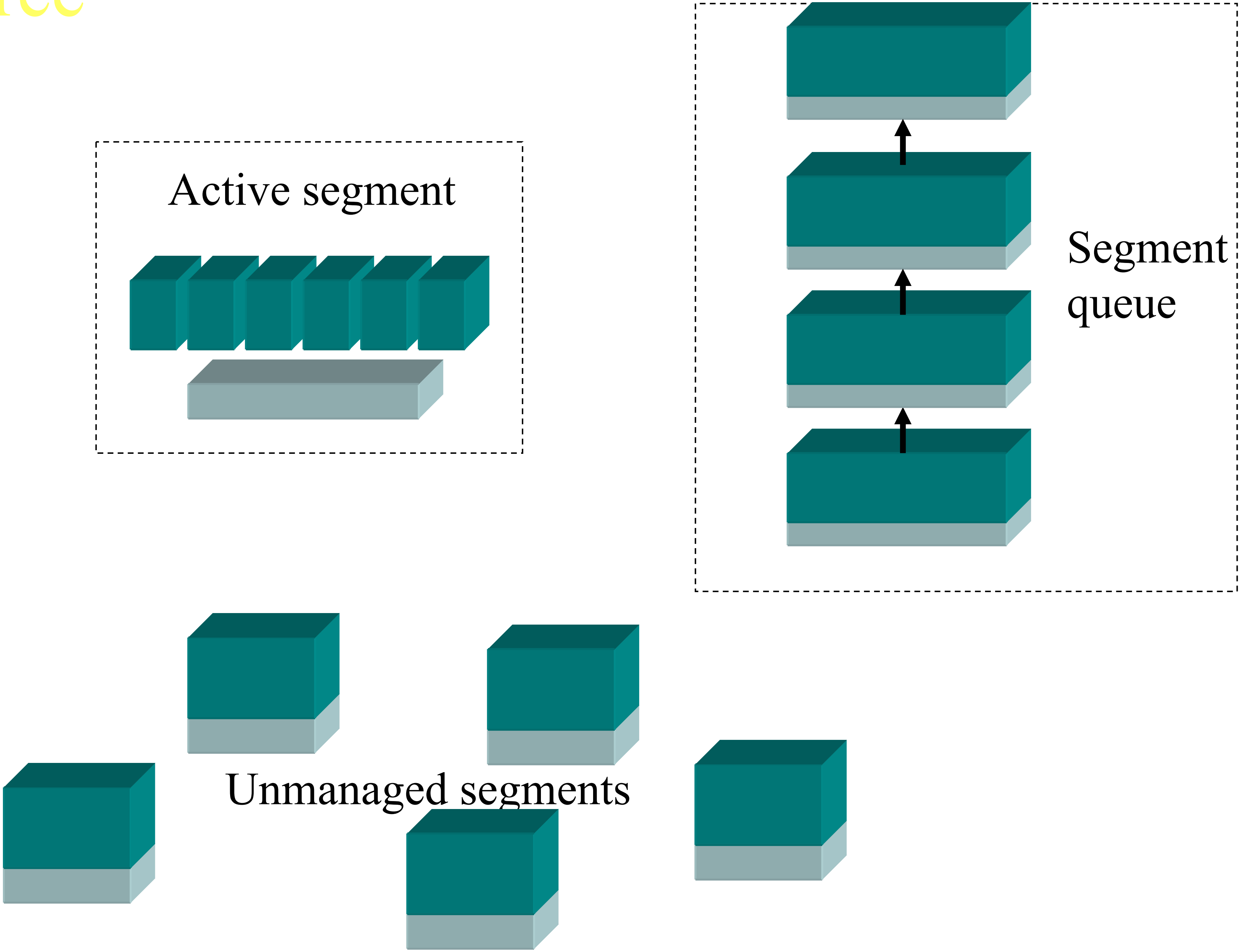
User data area

Descriptor

Segment queue

Unmanaged segments

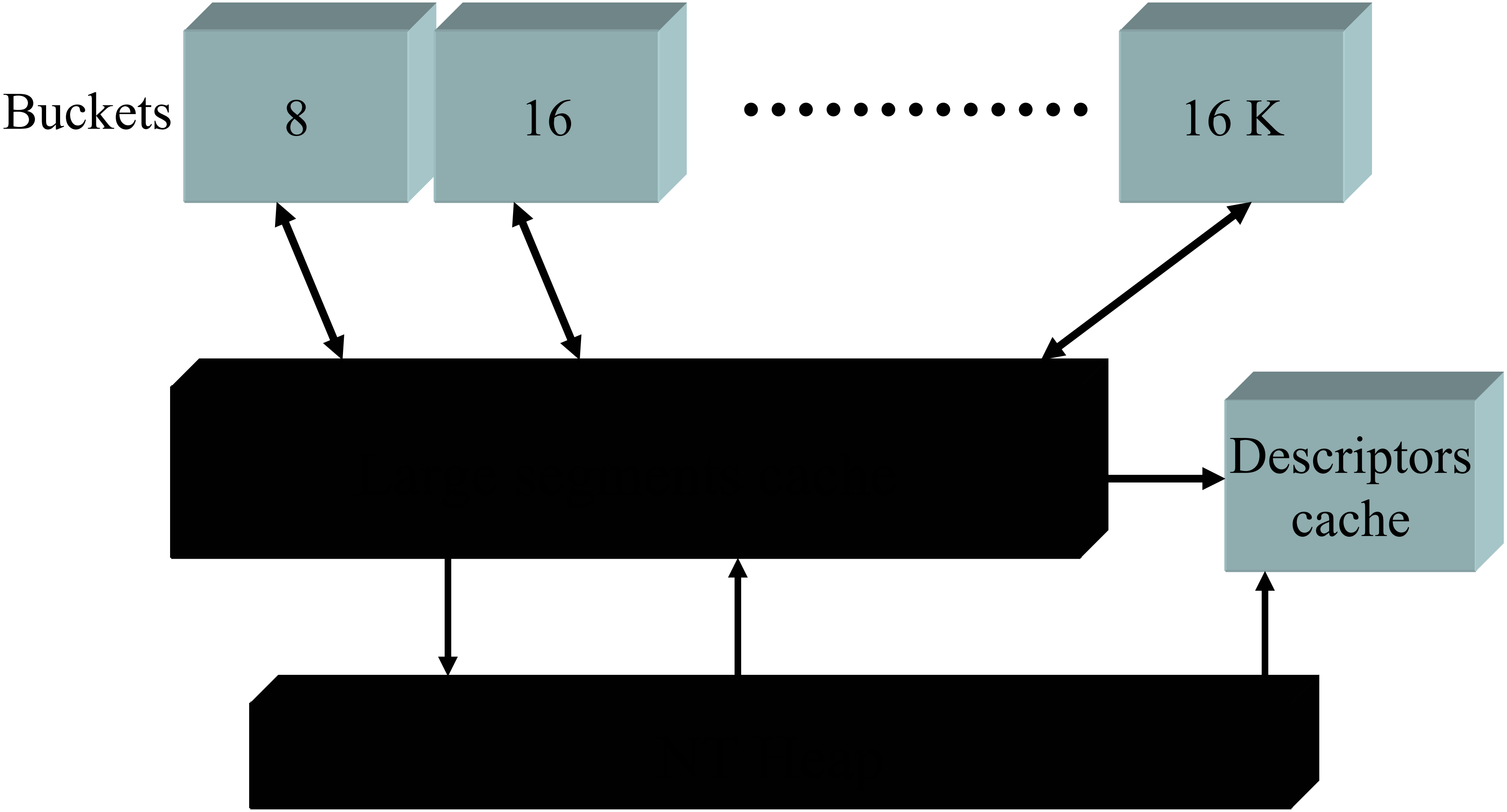
Free



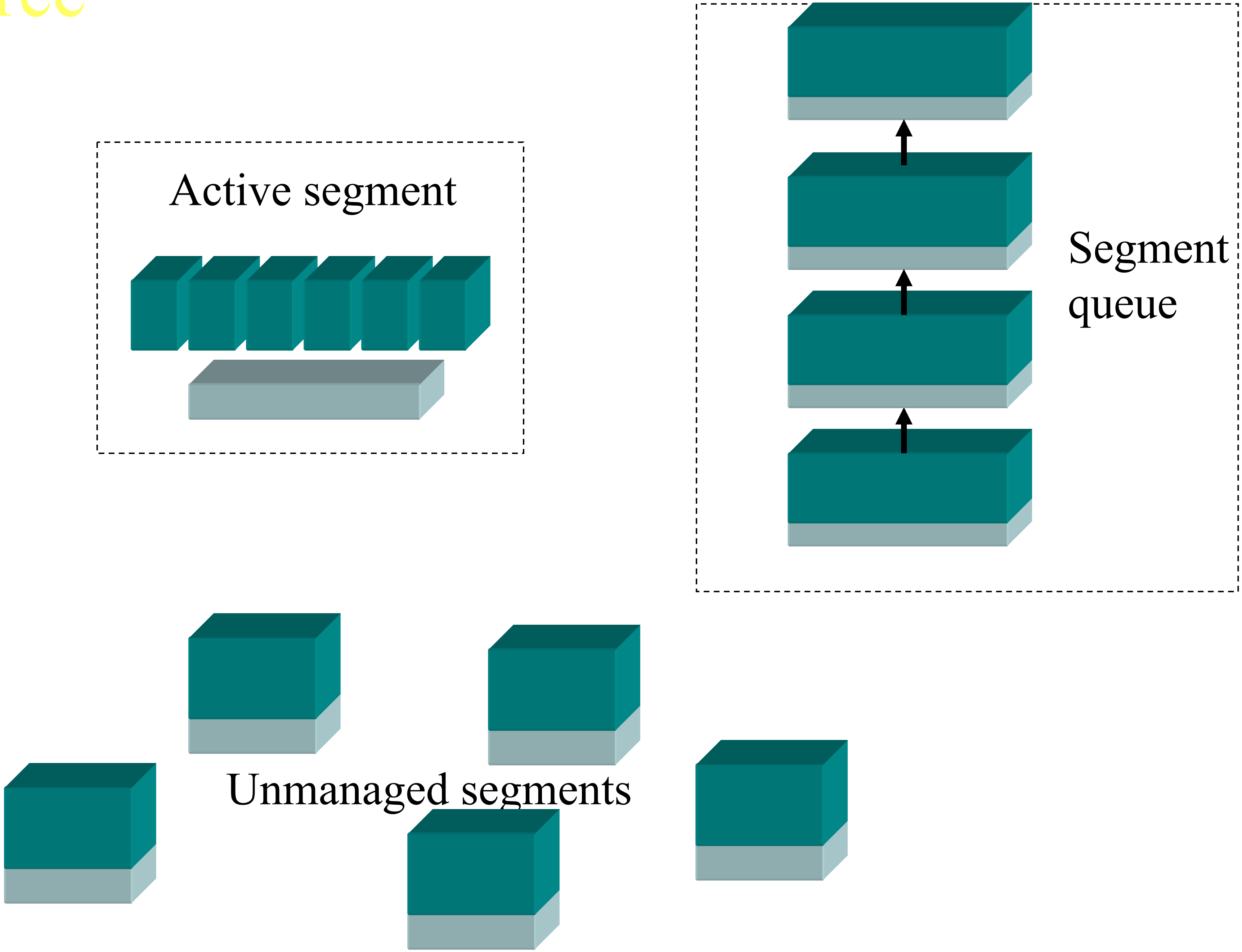
Active segment

Segment queue

Unmanaged segments


Buckets
8
16
16 K
Large segments cache
Descriptors cache
NT Heap

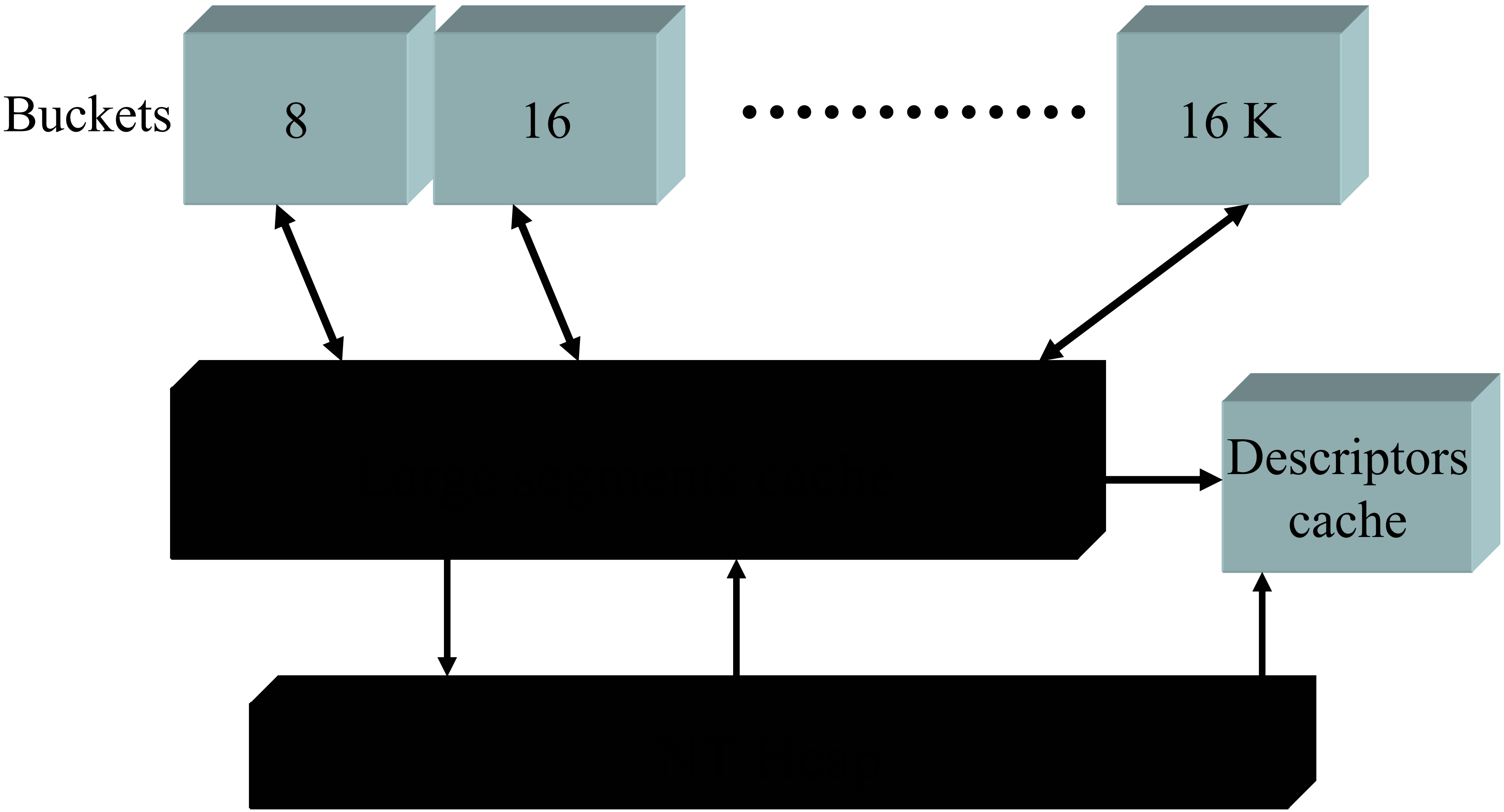
Free



Active segment

Segment queue

Unmanaged segments



Buckets
8
16
16 K
[redacted]
Descriptors cache
[redacted]

Improving the SMP Scalability

- Thread locality
- Processor locality

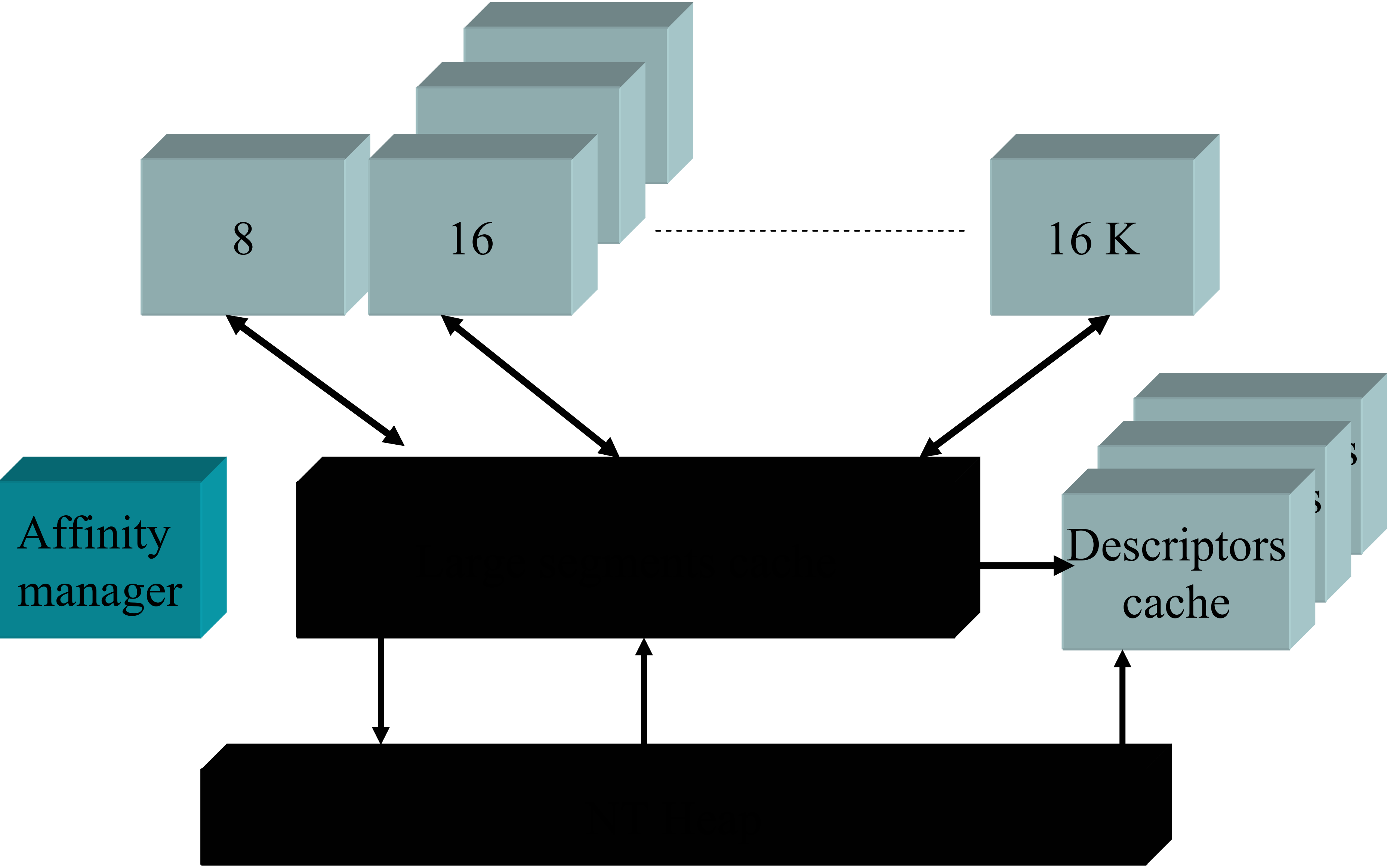
Thread Data Locality

- Advantages
  - Easy to implement (TLS)
  - Can reduce the number of interlocked instructions
- Disadvantages
  - Significantly larger footprint for high number of threads
  - Common source of leaks (the cleanup is not guaranteed)
  - Larger footprint for scenarios involving cross thread operations
  - Performance issues at low memory (larger footprint can cause paging)
  - Increases the CPU cost per thread creation / deletion

Processor Locality

- Advantages
  - The memory footprint is bounded to the number of CPUs regardless of the number of threads
  - Expands the structures only if needed
  - No cleanup issues
- Disadvantages
  - The current CPU is not available in user mode
  - Not efficient for a large number of processors and few threads

MP Scalability



Better Than Lookaside

- Better data locality (likely in same page)
- Almost perfect SMP scalability (no false sharing)
- Covers a larger size range (up to 16k blocks)
- Works well regardless of the number of blocks
- Non-blocking operations even during growing and shrinking phases

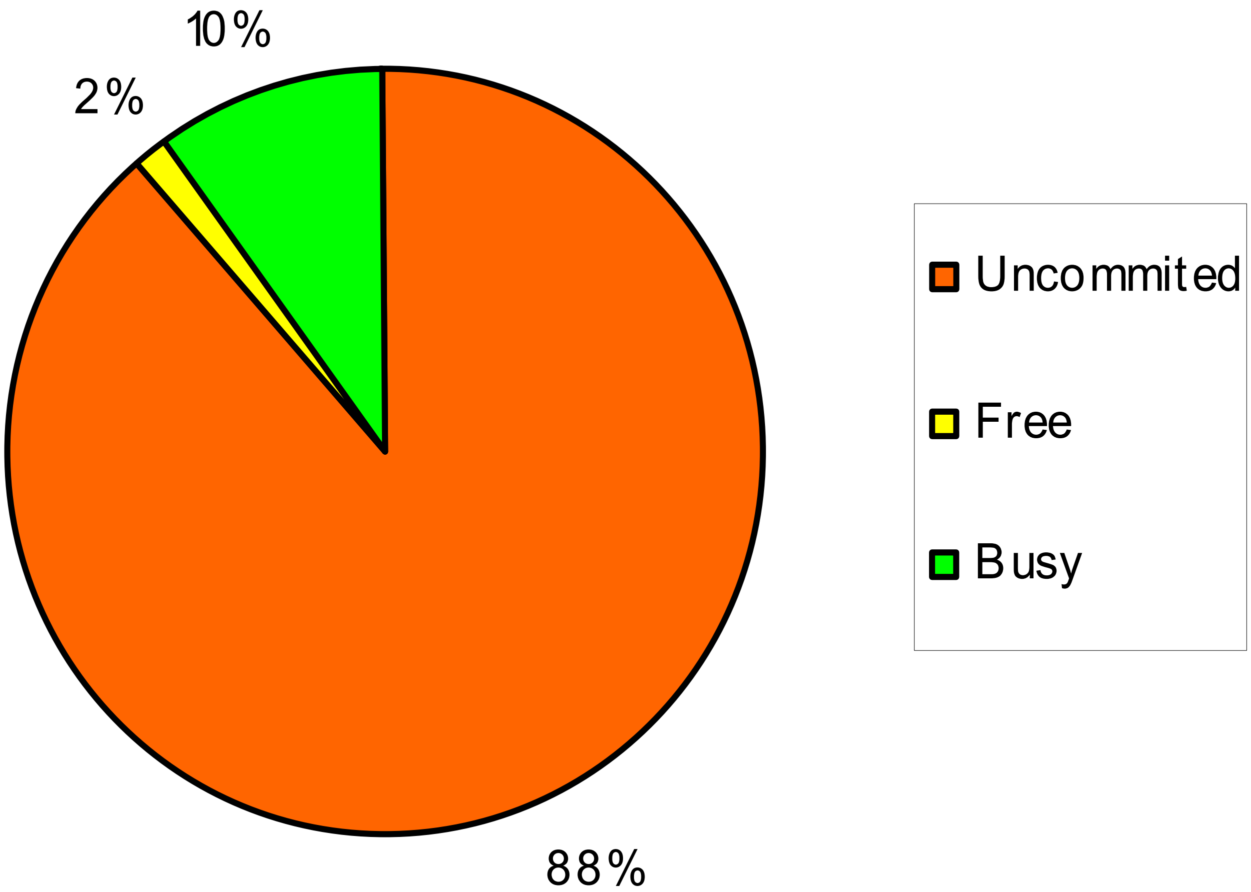
Benchmarks

- **Fragmentation**
- **Speed**
- **Scalability**
- **Memory efficiency**

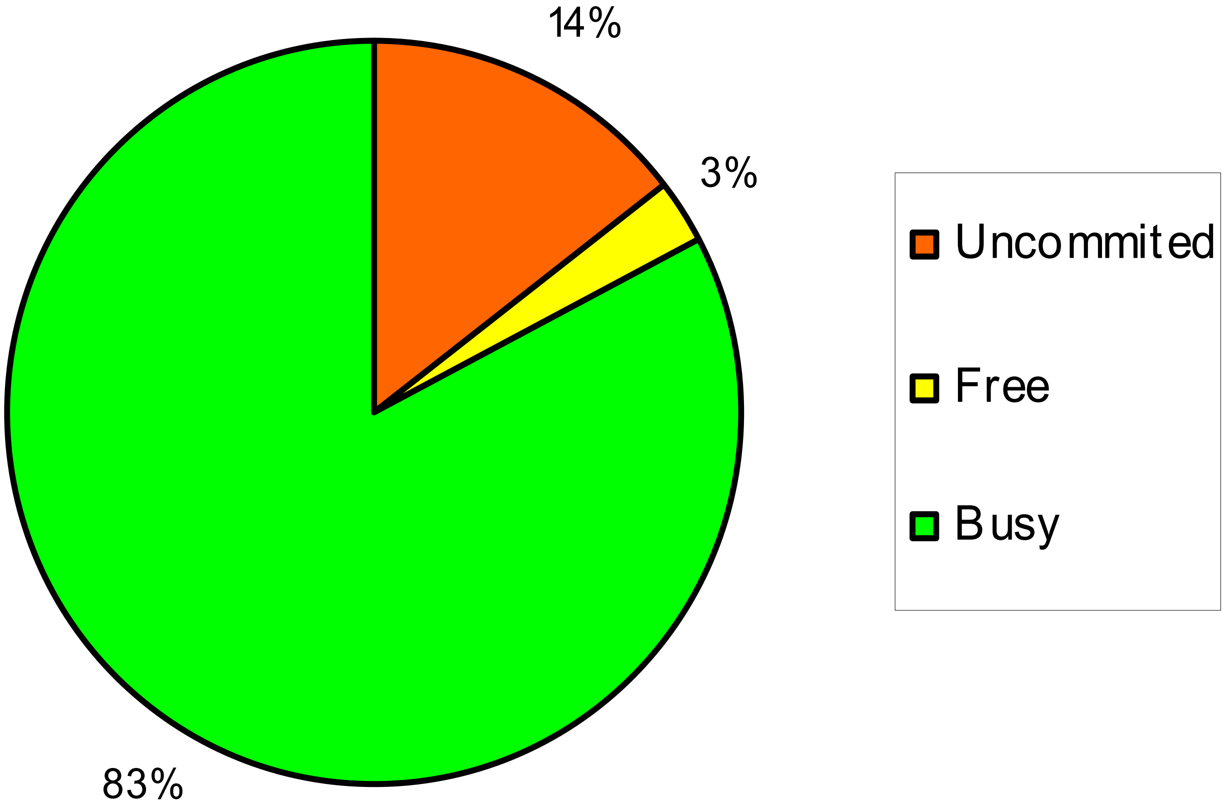
Fragmentation
test for 266 MB limit

	Default	LFH
Uncommited	235 M**B**	39 MB
Free	4 MB	7 MB
Busy	26 MB	224 MB
Fragmentation	**88%**	**14%**

Default NT Heap



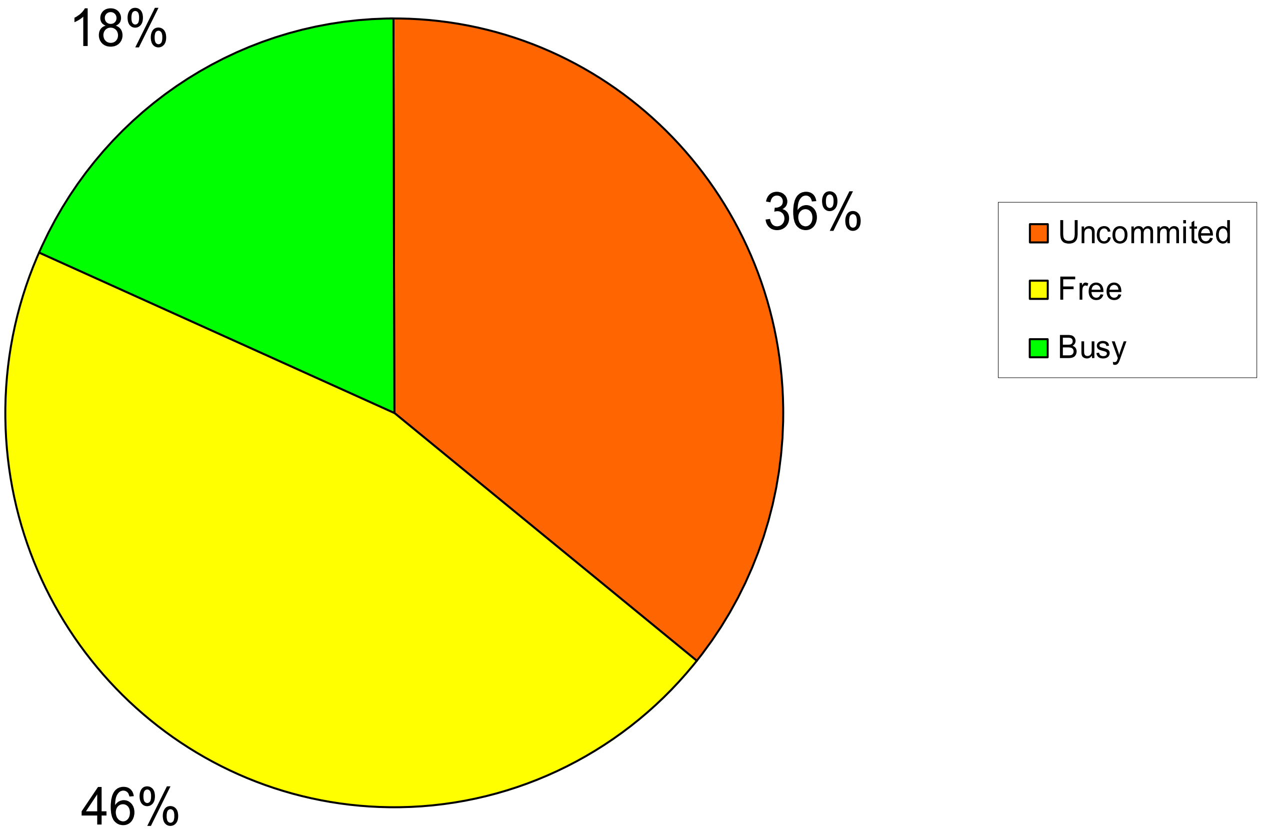
Low Fragmentation Heap



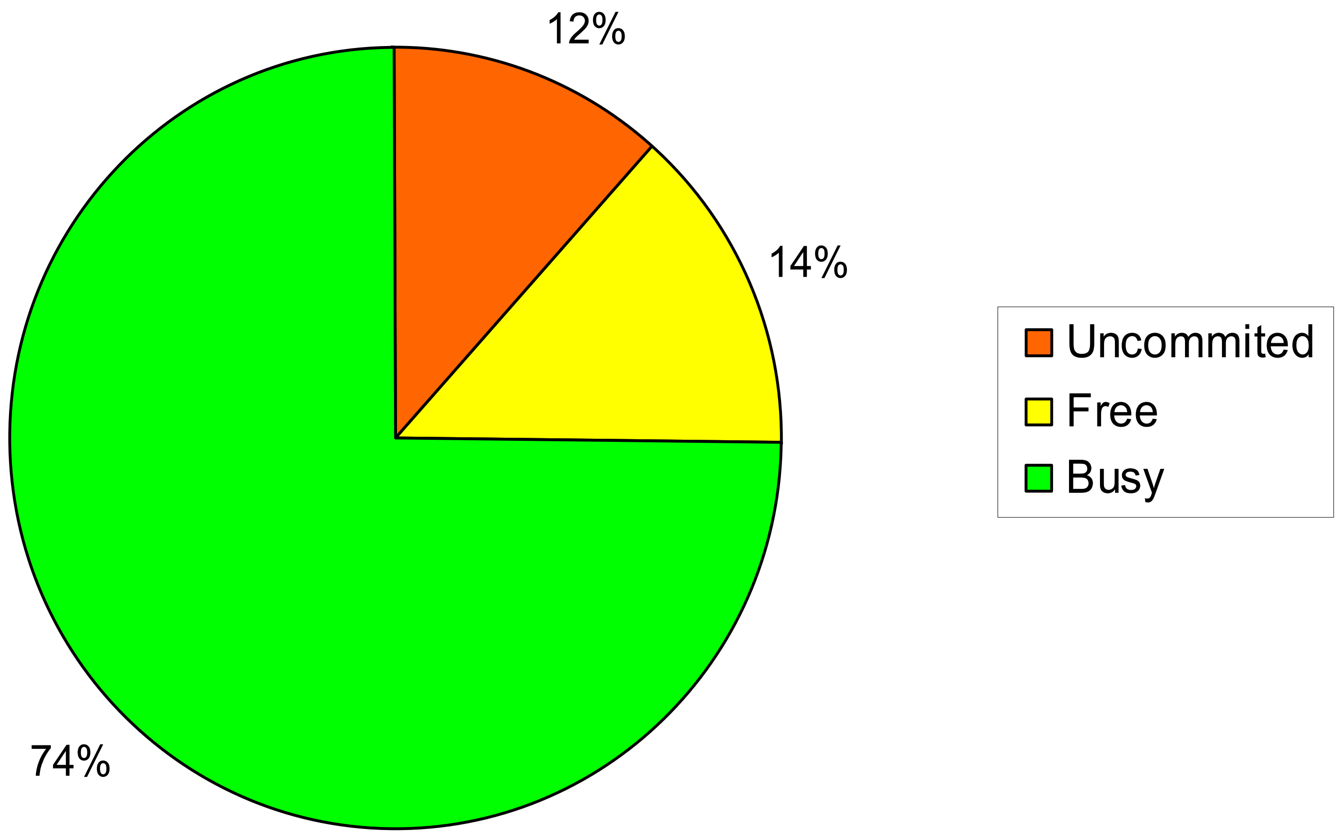
External Fragmentation Test (70 MB)

	Default	LFH
Uncommited	25 MB	7 MB
Free	32 MB	8 MB
Busy	12 MB	46 MB
Fragmentation	**46% + 36%**	**14% + 12%**

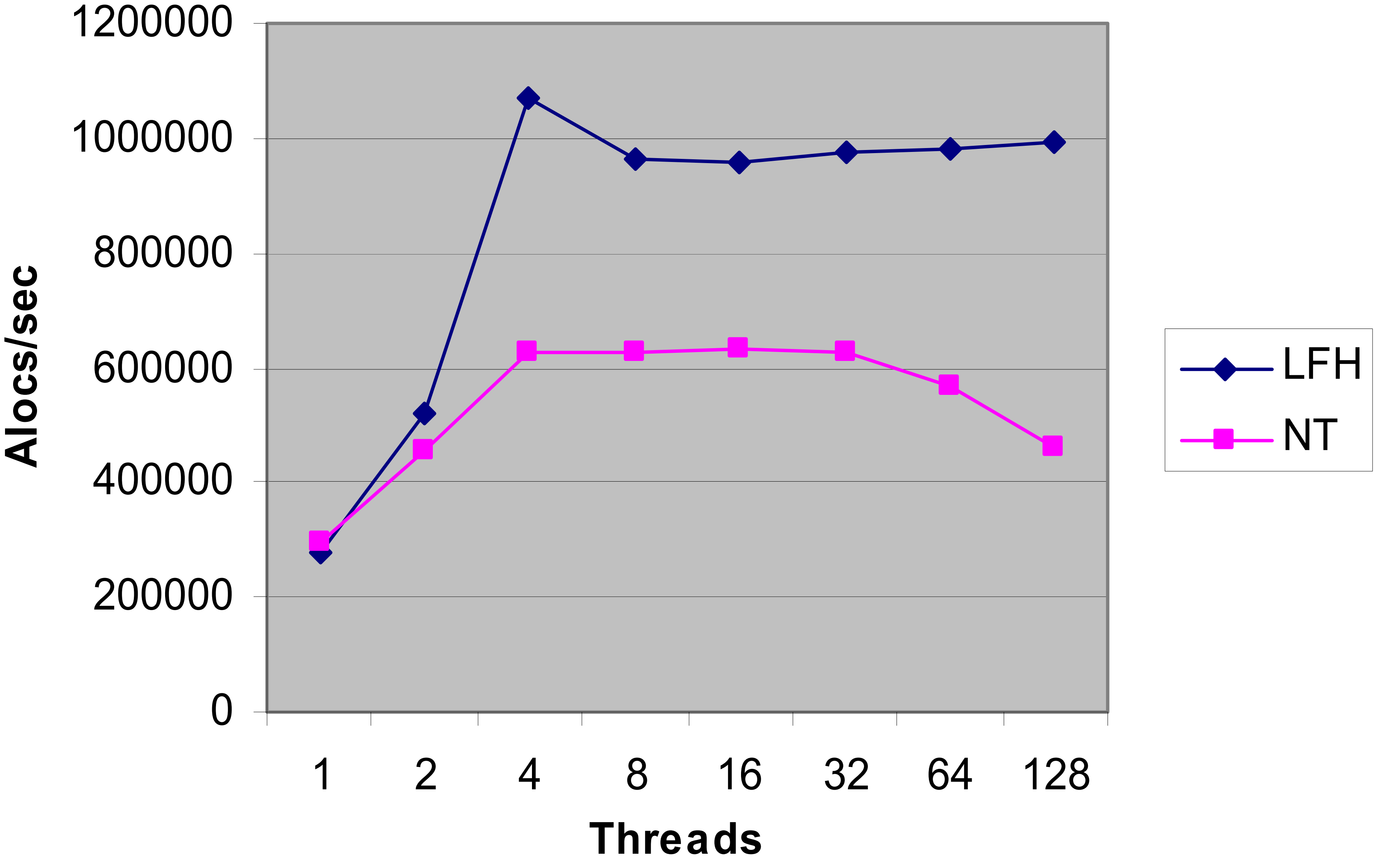
NT Heap at 70 M usage
(8478 UCR, 10828 free blocks)



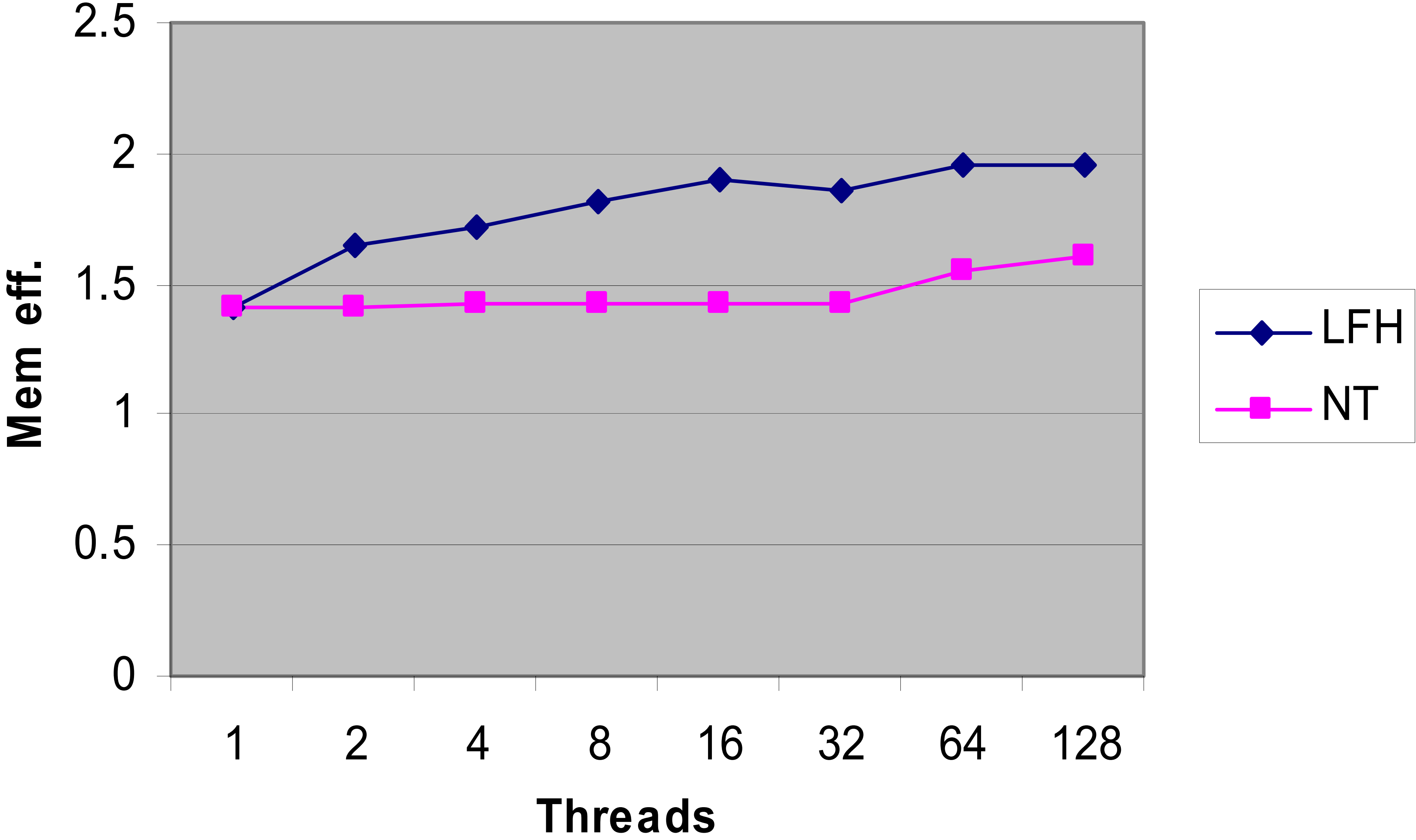
Low Fragmentation Heap at 70 M
(417 UCR, 1666 free blocks)



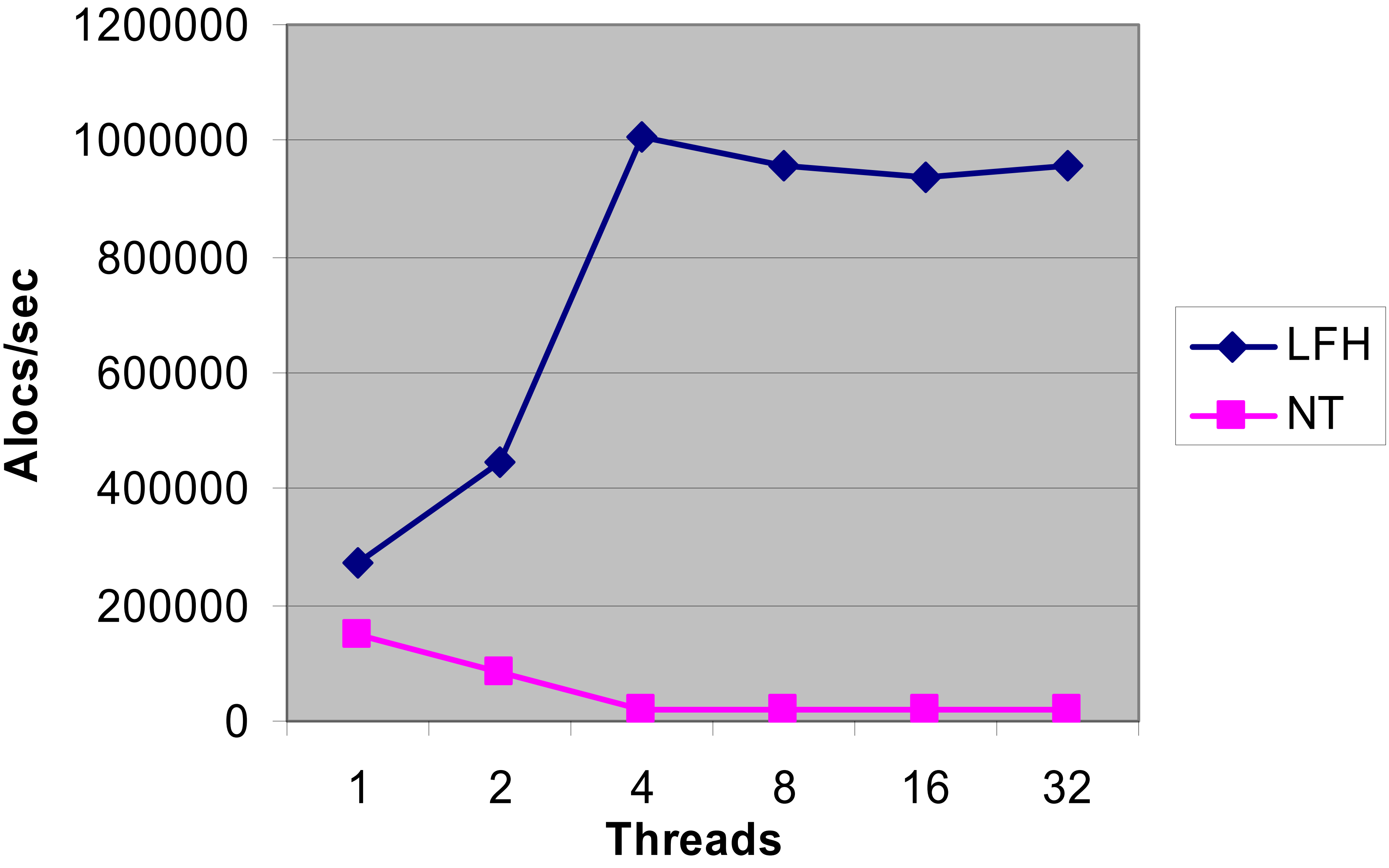
Replacement test
0-1k, 10000 blocks (4P x 200MHz)



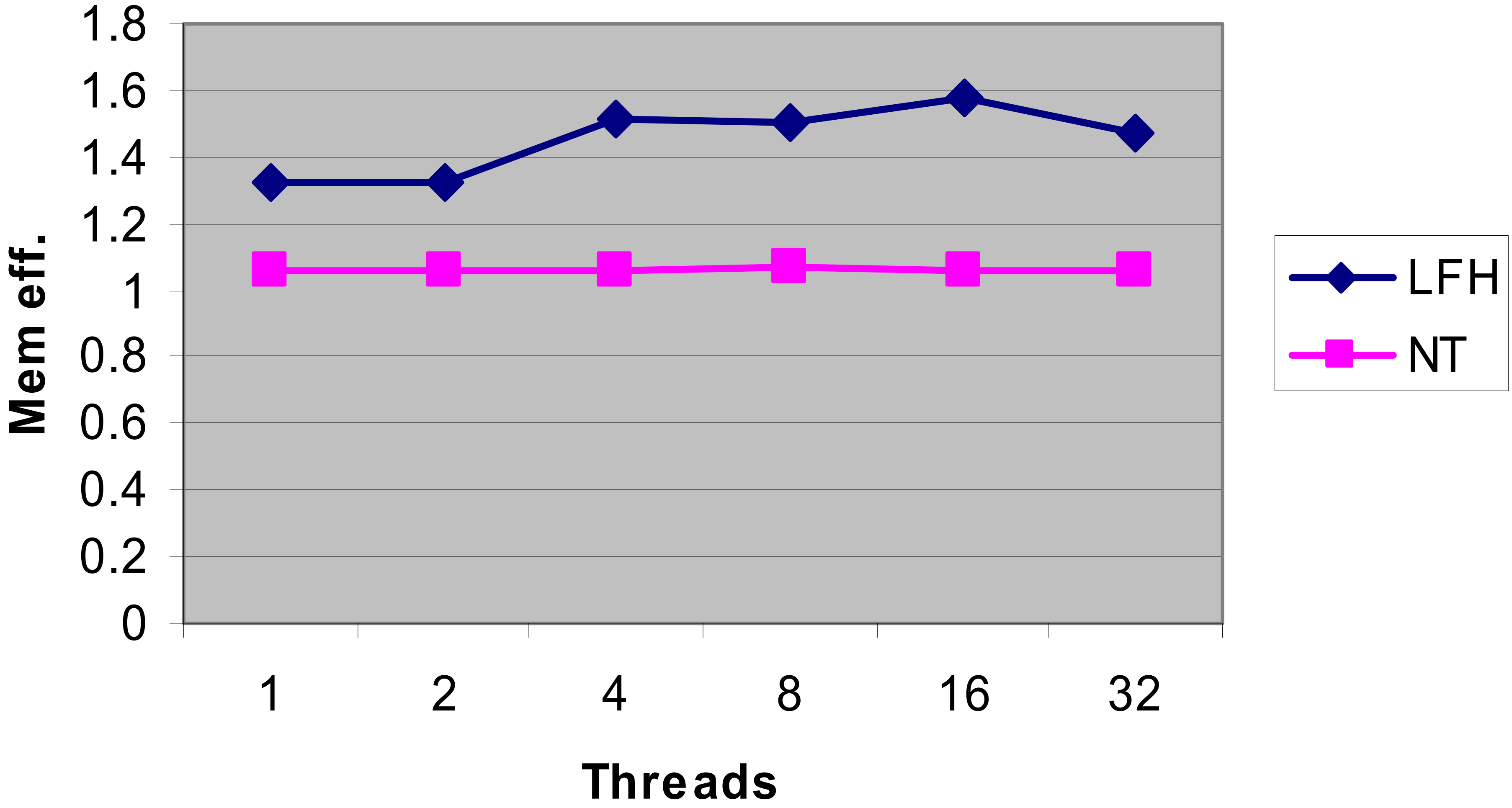
Replacement test
0-1k, 10000 blocks



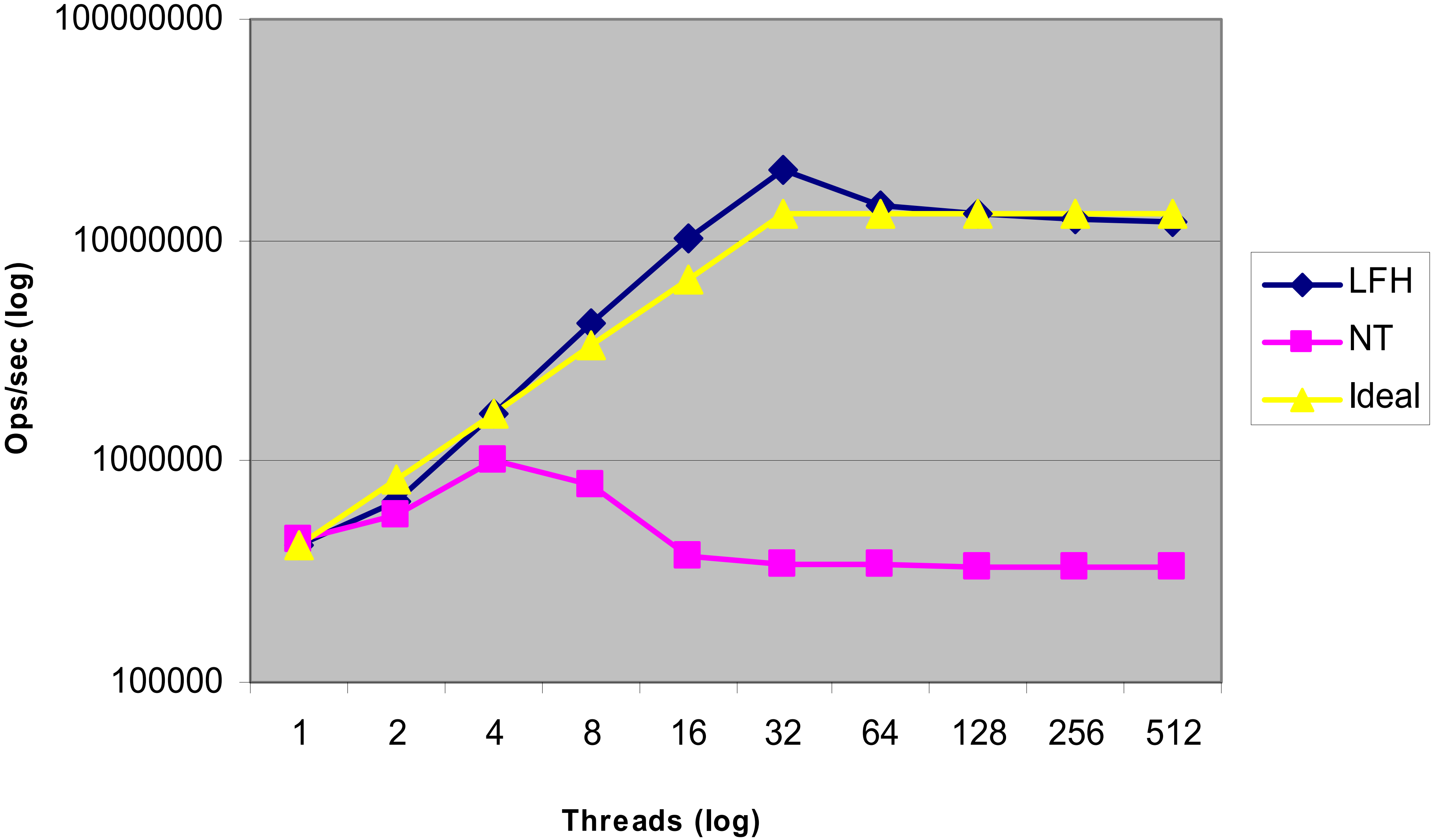
Replacement test
1-2k, 10000 blocks



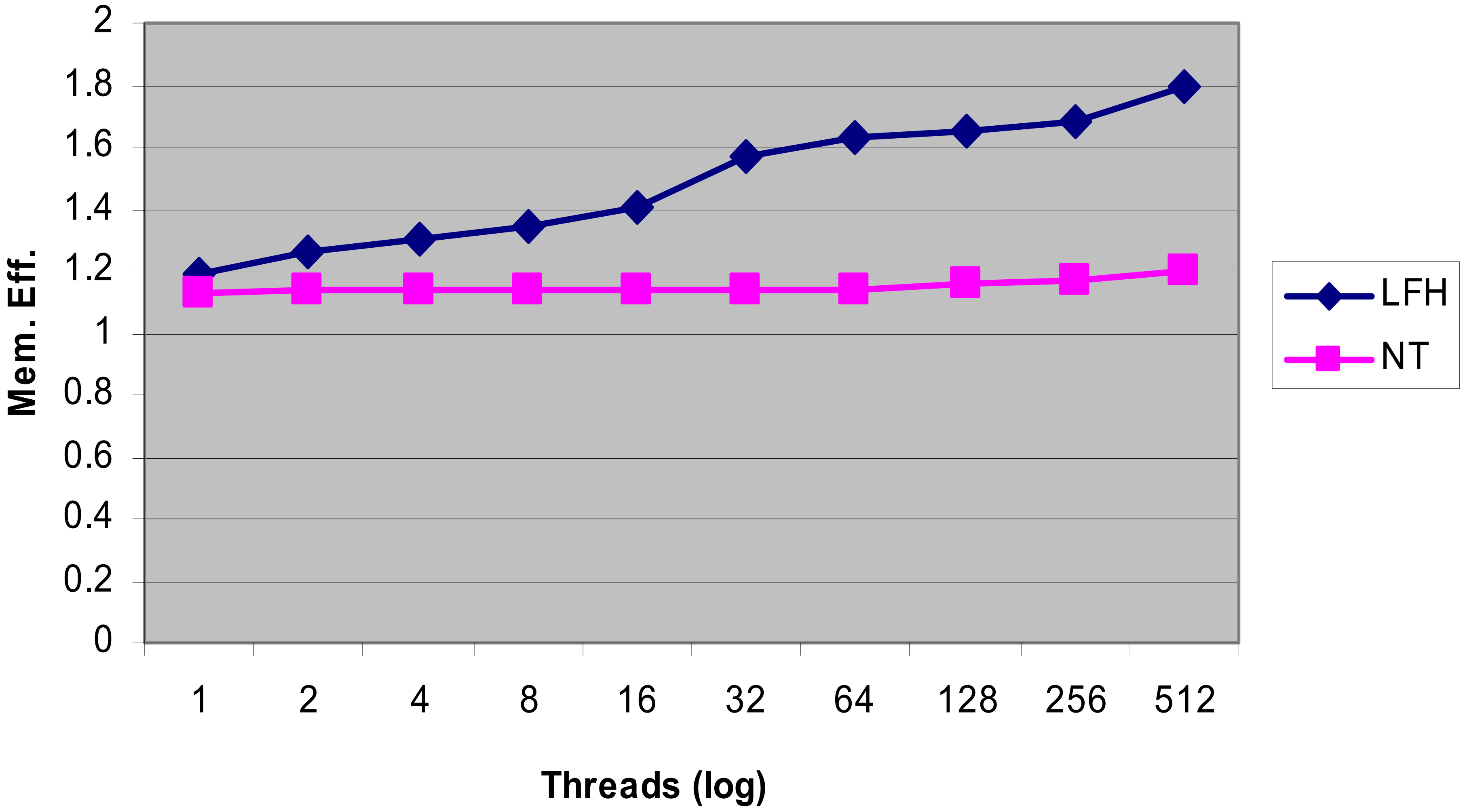
Replacement test
1-2k, 10000 blocks



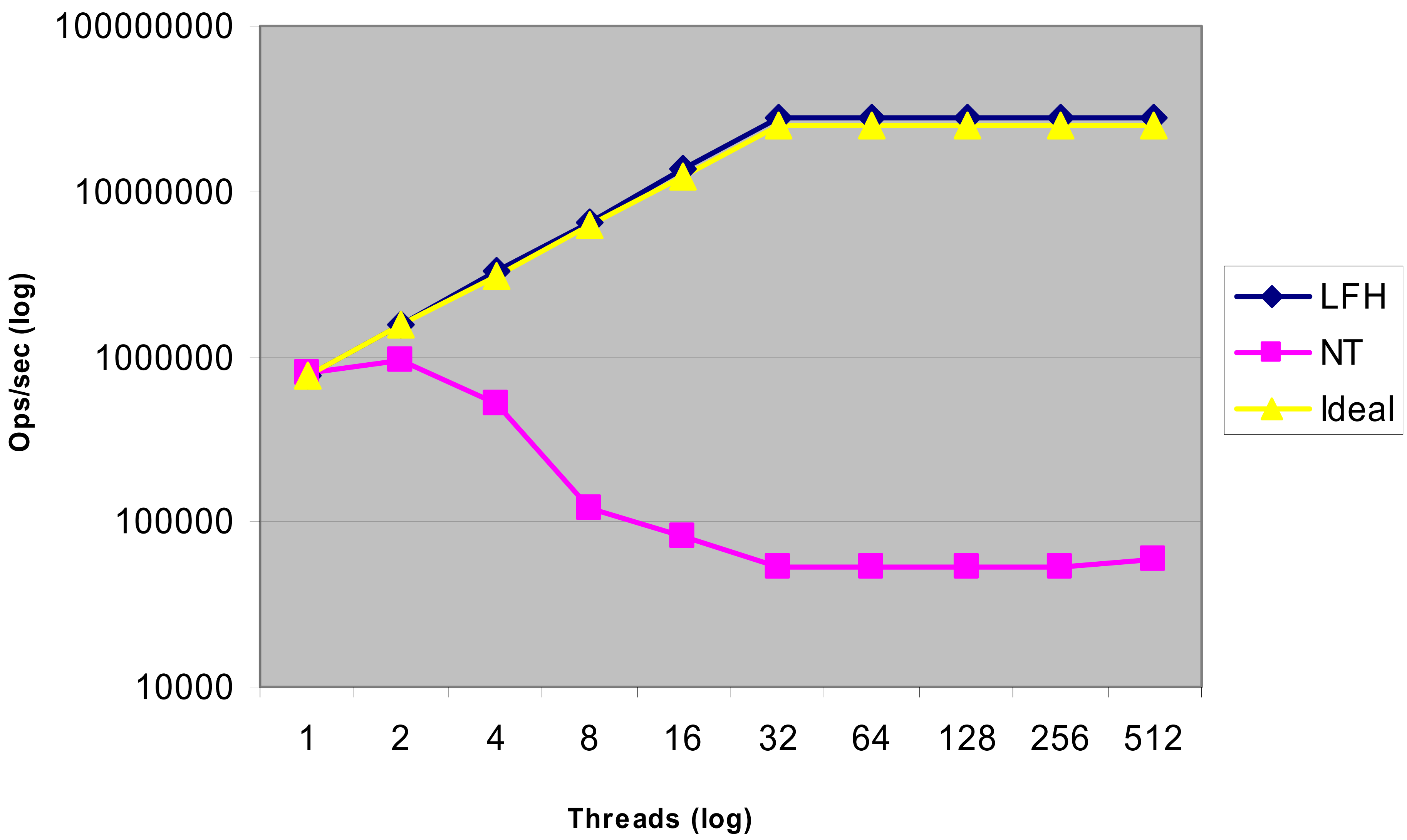
Replacement test on a 32P machine
0-1k, 100000 blocks



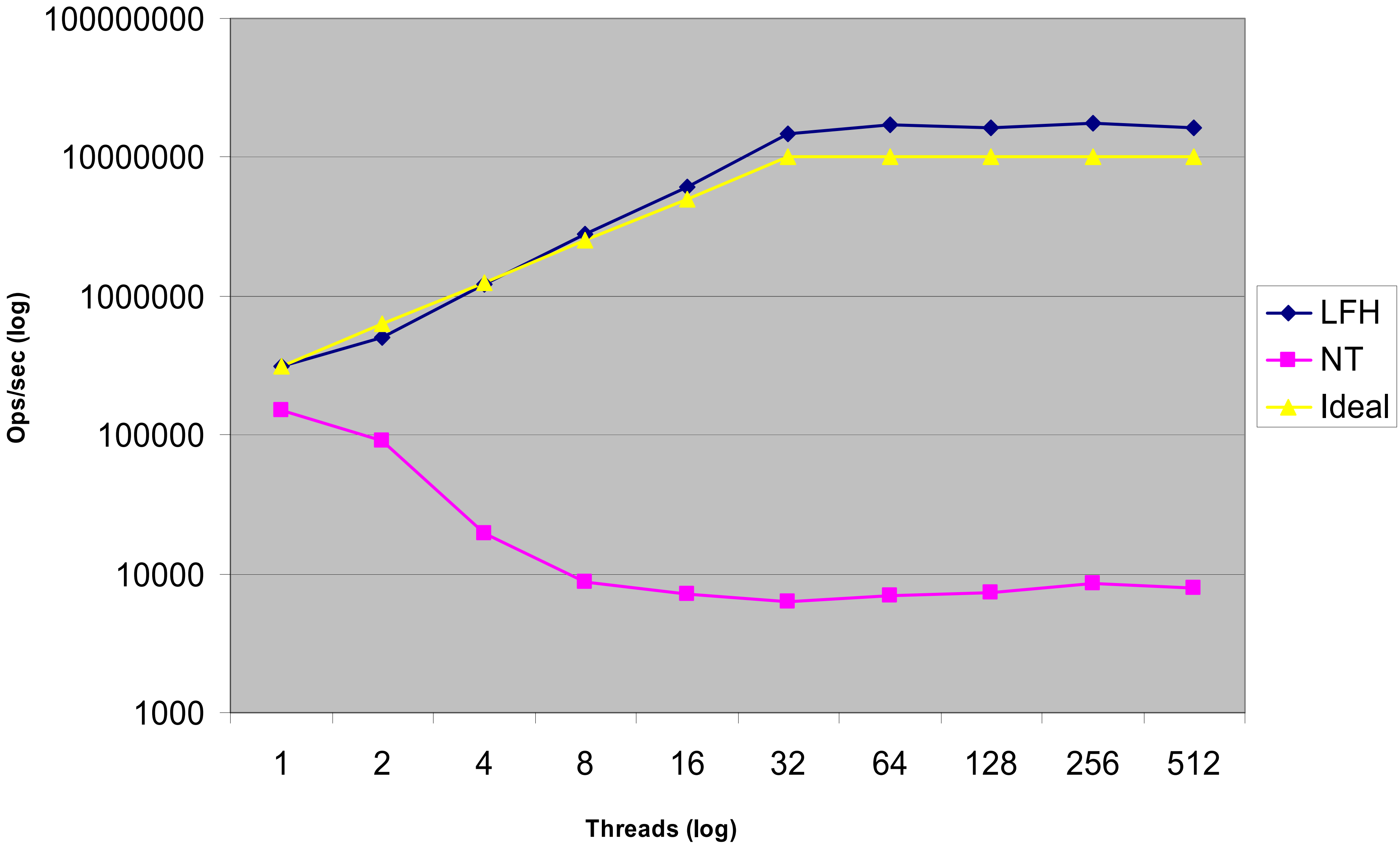
Replacement test on 32P machine
0-1k, 100000 blocks



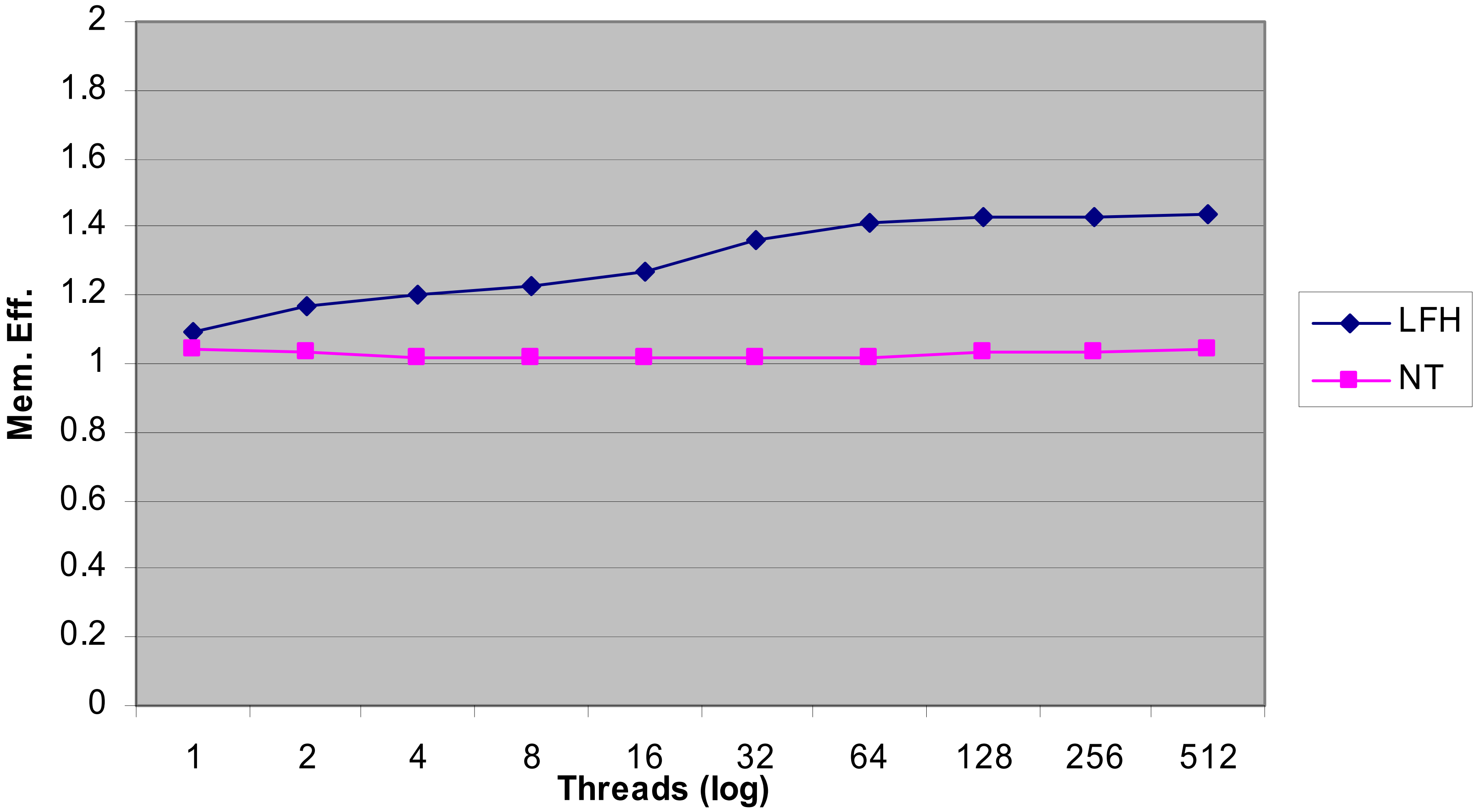
Replacement test on 32P machine
22 bytes, 100000 blocks



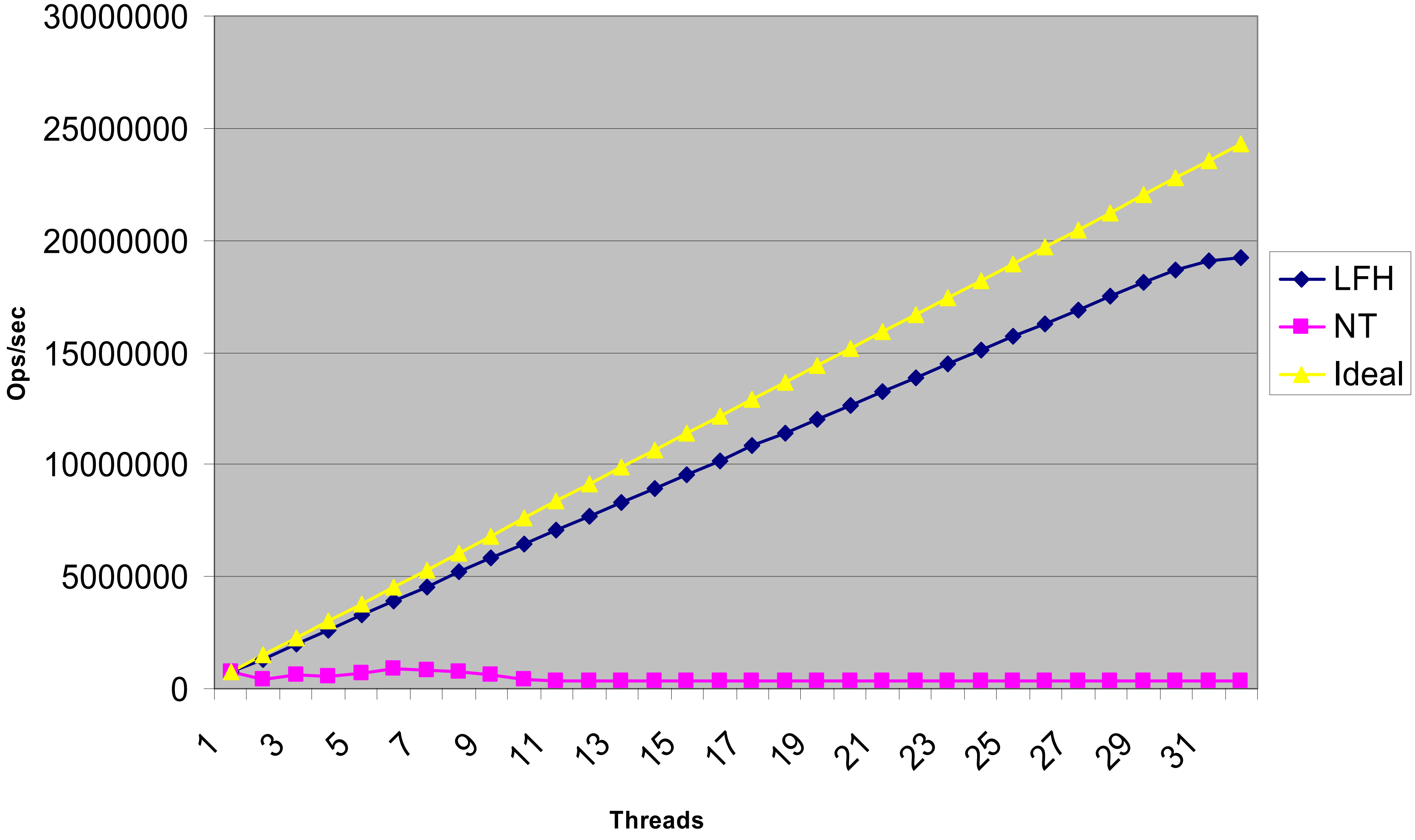
Replacement test on 32P machine
1k-2k, 100000 blocks



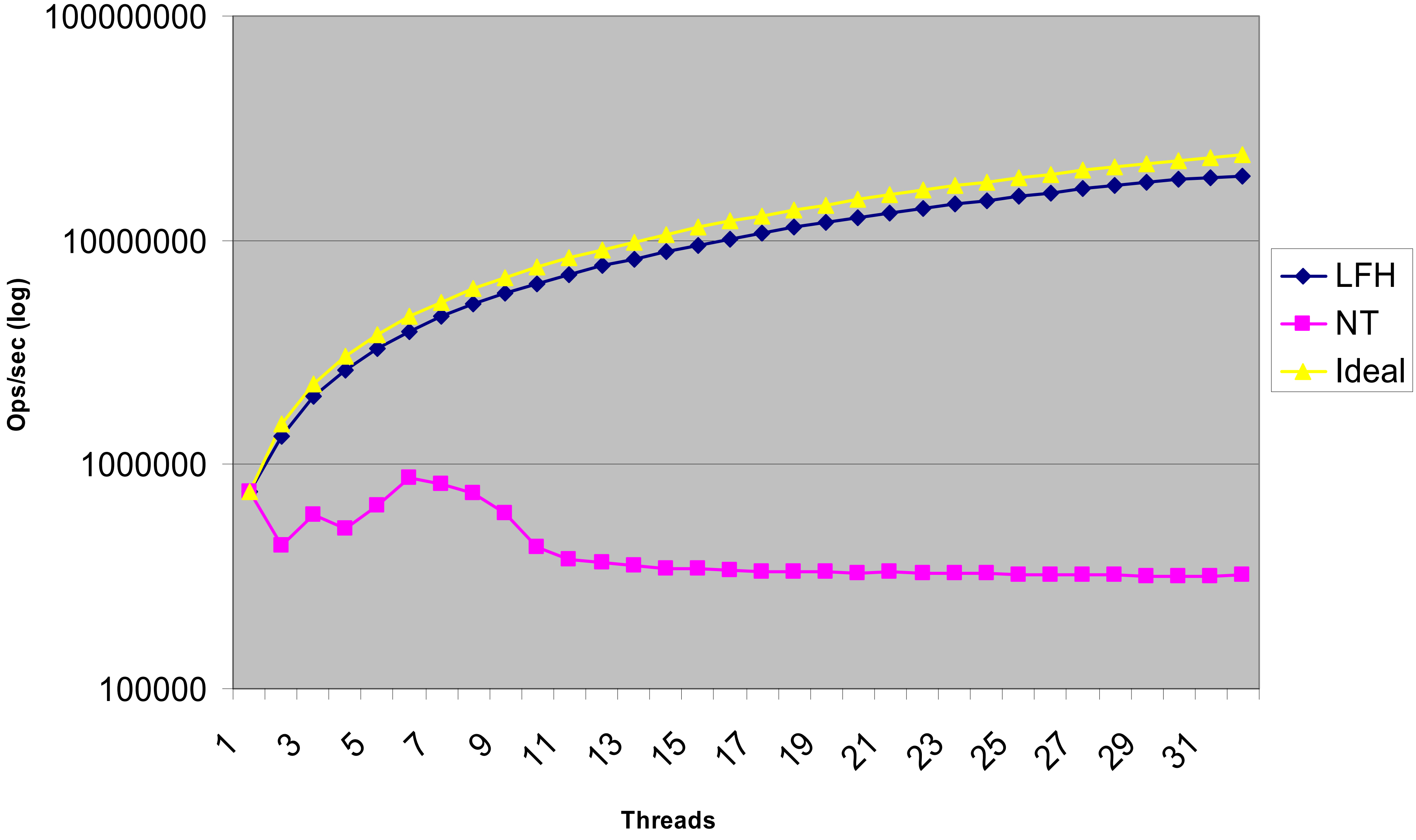
Replacement test on 32P machine
1k-2k, 100000 blocks



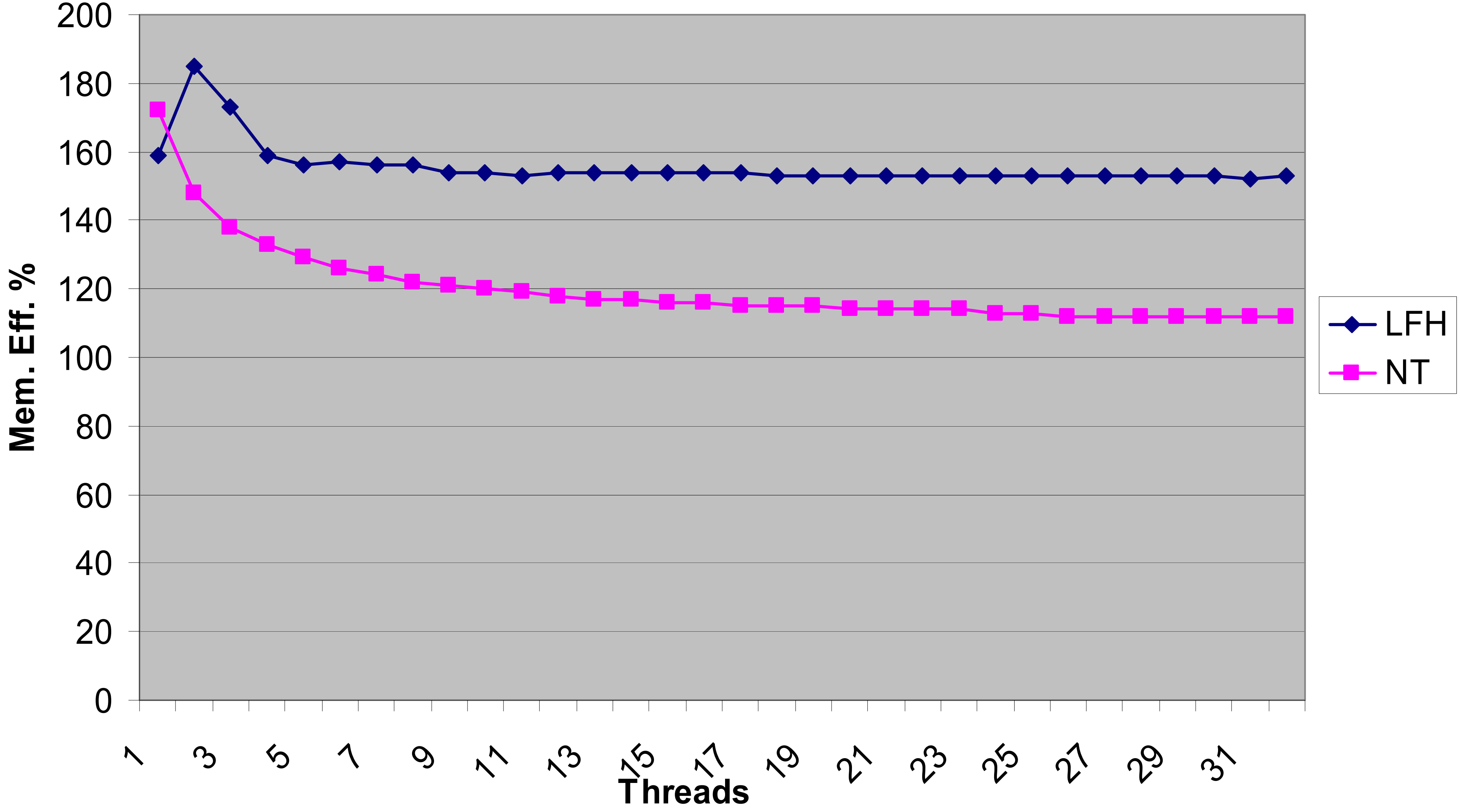
Larson MT test on 32P machine
0 - 1k, 3000 blocks/thread



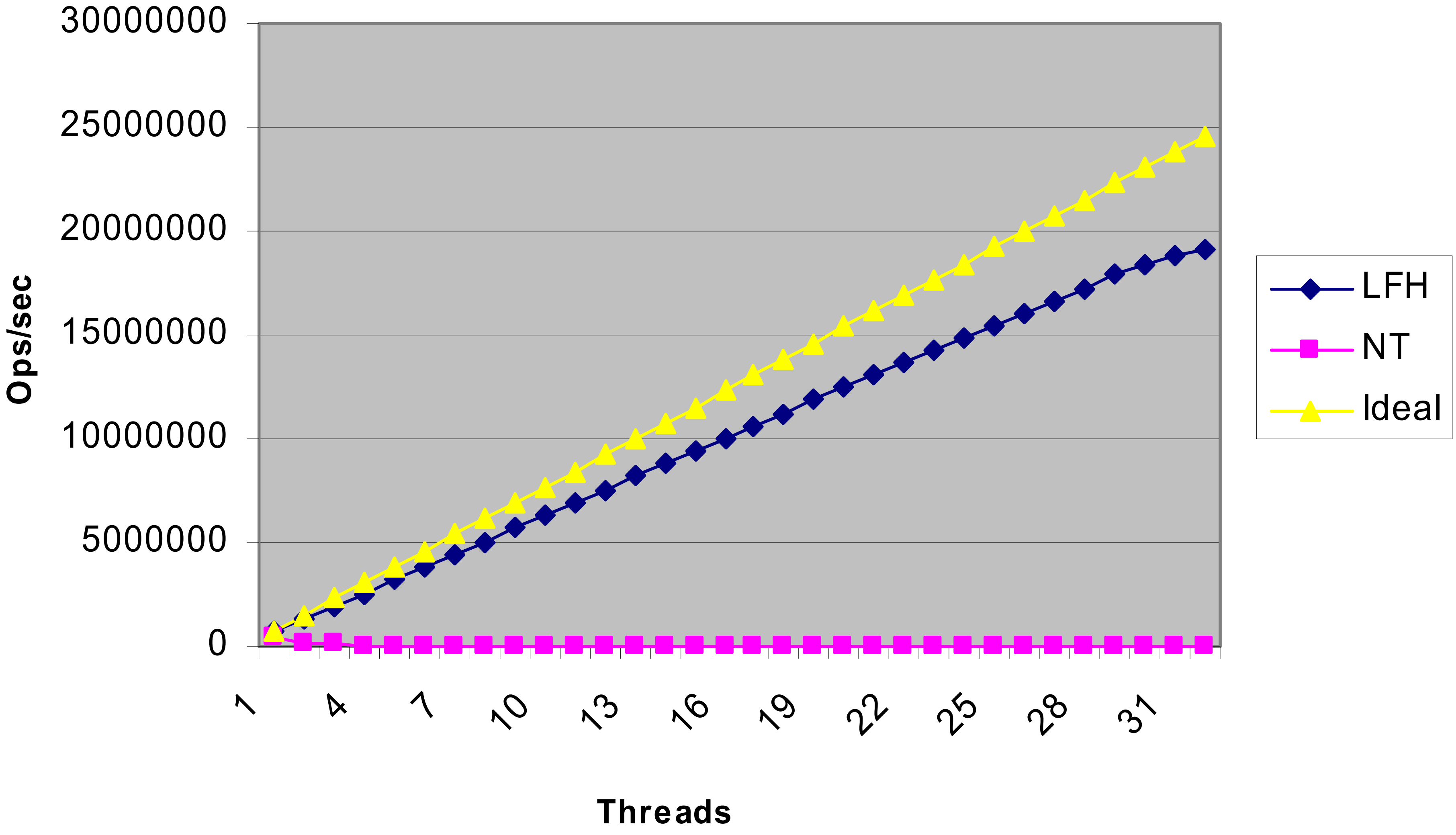
Larson MT test on 32P machine
0 - 1k, 3000 blocks/thread



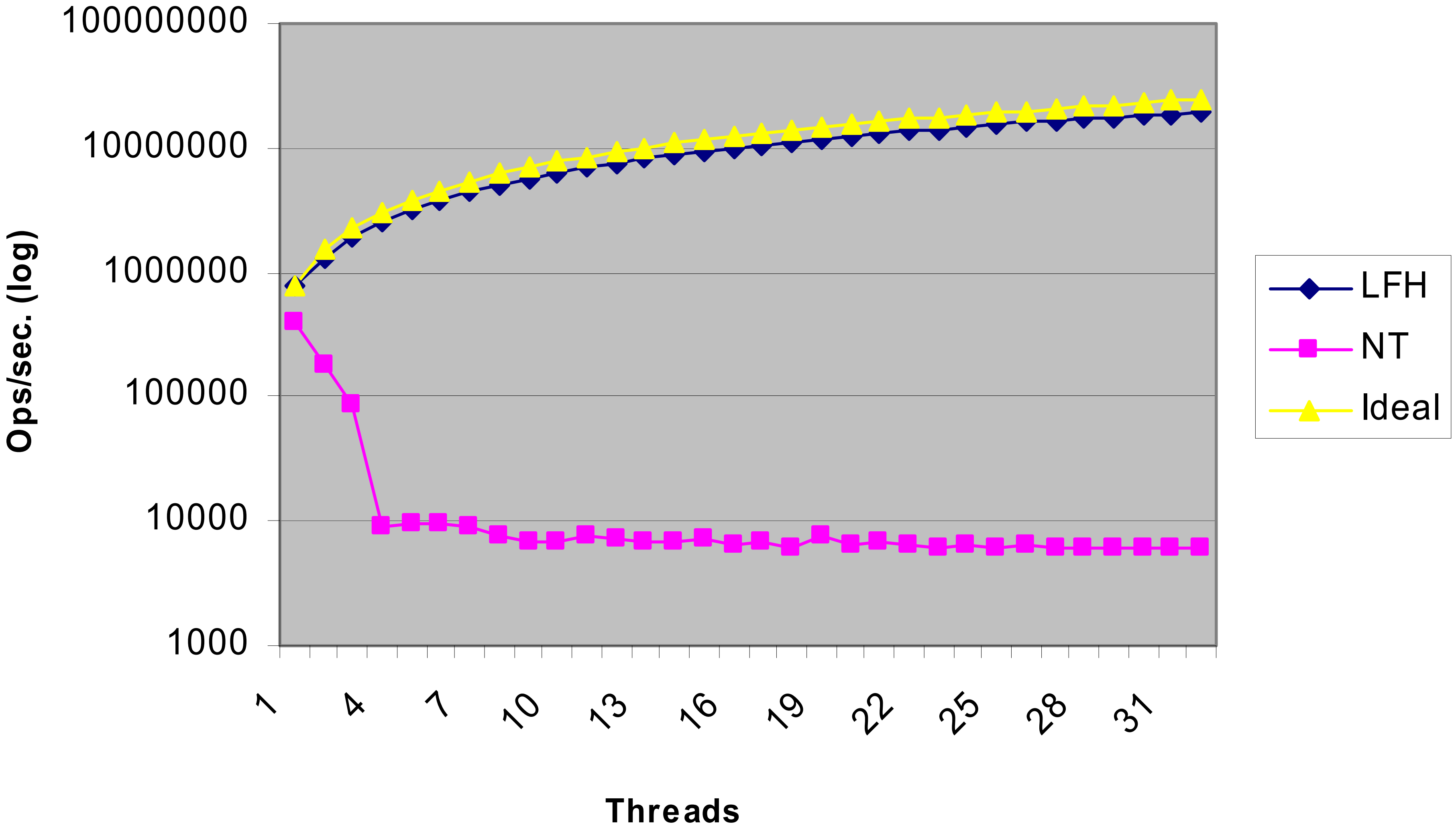
Larson MT test on 32P machine
0 - 1k, 3000 blocks / thread



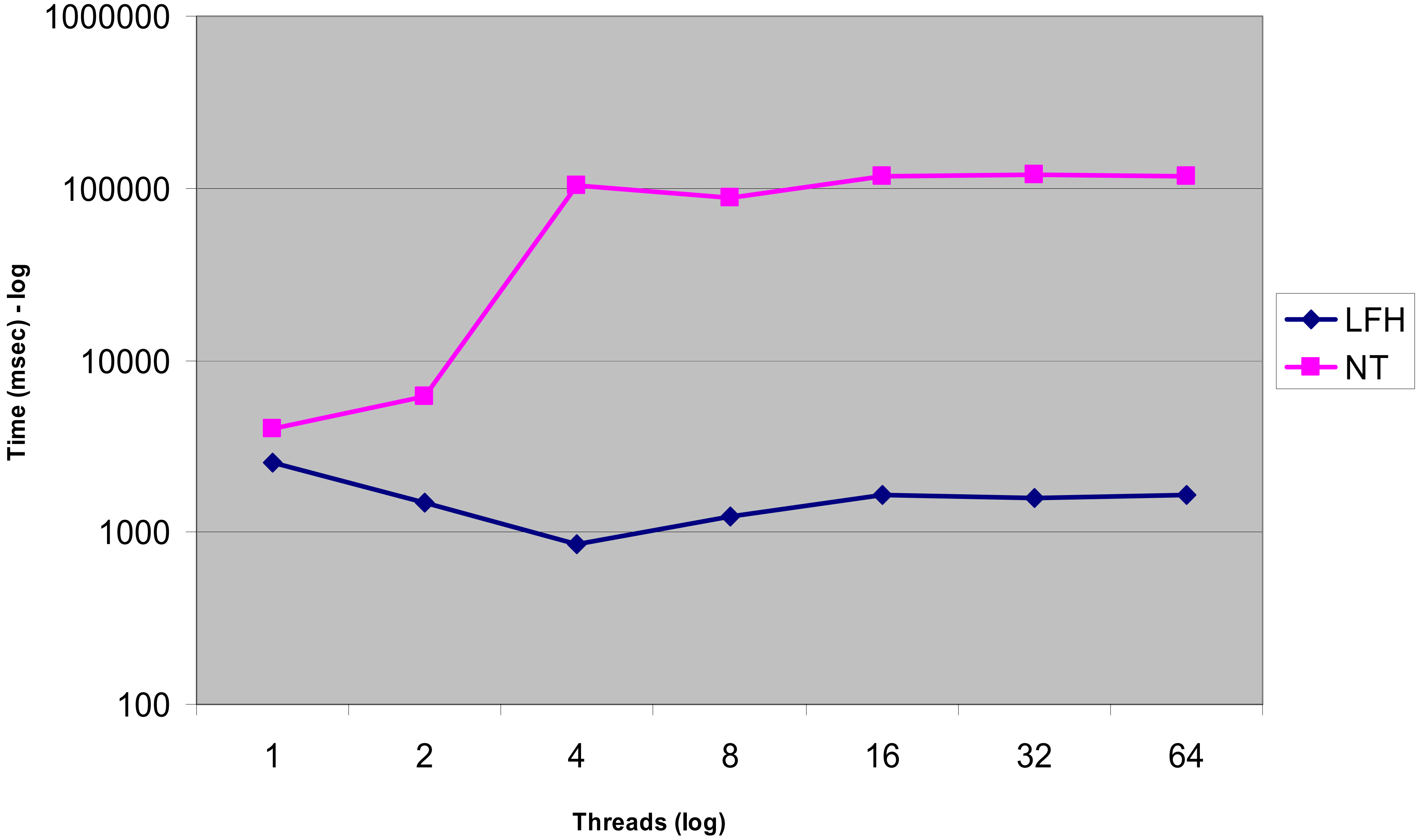
Larson MT test on 32P machine
1k -2k, 100000 blocks



Larson MT test on 32P machine
1k -2k, 100000 blocks



Aggressive alloc test on 32P machine
50 Mbytes allocs in blocks of 32 bytes



When is the Default Heap Preferred

- **~95% of applications**
- **The heap operations are rare**
- **Low memory usage**

Where LFH is Recommended

- High memory usage and:
  - High external fragmentation (> 10-15%)
  - High virtual address fragmentation (>10-15%)
- Performance degradation on long run
- High heap lock contention
- Aggressive usage of large blocks (> 1K)

Activating LFH

- HeapSetInformation
  - Can be called any time after the heap creation
  - Restriction for some flags (HEAP_NO_SERIALIZE, debug flags)
  - Can be destroyed only with the entire heap

- HeapQueryInformation
  - Retrieve the current front end heap type
    - 0 – none
    - 1 – lookaside
    - 2 – LFH

Heap Analysis

- !heap to collect statistics and validate the heap
  - **!heap –s**
  - **!heap –s** *heap_addr* **–b***8*
  - **!heap –s** *heap_addr* **–d***40*
- Perfmon

Overall Heap Stats

```
0:001> !heap -s

  Heap     Flags   Reserv  Commit  Virt   Free  List   UCR  Virt  Lock  Fast
                    (k)     (k)    (k)     (k) length      blocks cont. heap
-----------------------------------------------------------------------------
00080000 00000002    1024      28     28     14     1     1    0      0   L
00180000 00008000      64       4      4      2     1     1    0      0
00250000 00001002      64      24     24      6     1     1    0      0   L
00270000 00001002  130304  58244  96888  36722 10828  8478    0      0   L
    External fragmentation  63 % (10828 free blocks)
    Virtual address fragmentation  39 % (8478 uncommited ranges)
-----------------------------------------------------------------------------
```

Overall Heap Stats

```
0:000> !heap -s

  Heap     Flags   Reserv  Commit  Virt   Free  List   UCR  Virt  Lock  Fast
                    (k)     (k)    (k)     (k) length      blocks cont. heap
-----------------------------------------------------------------------------
00080000 00000002    1024     28     28     16     2     1    0      0
00180000 00008000      64      4      4      2     1     1    0      0
00250000 00001002      64     24     24      6     1     1    0      0
00270000 00001002     256    116    116      5     1     1    0      0
002b0000 00001002  130304 122972 122972   1936    67     1    0 14d5b8
    Lock contention     1365432
-----------------------------------------------------------------------------
```

Overall Heap Stats

```
0:006> !heap -s

The process has the following heap extended settings 00000008:
   - Low Fragmentation Heap activated for all heaps

Affinity manager status:
   - Virtual affinity limit 8
   - Current entries in use 4
   - Statistics:  Swaps=18, Resets=0, Allocs=18

  Heap     Flags   Reserv  Commit  Virt   Free  List   UCR  Virt  Lock  Fast
                    (k)     (k)    (k)     (k) length      blocks cont. heap
-----------------------------------------------------------------------------
00080000 00000002    1024    432    432      2     1     1    0      0   LFH
00180000 00008000      64      4      4      2     1     1    0      0
00250000 00001002    1088    364    364      1     1     1    0      0   LFH
00370000 00001002     256    212    212      3     1     1    0      0   LFH
003b0000 00001002    7424   5720   6240     43     3    26    0      f   LFH
-----------------------------------------------------------------------------
```

Default NT Heap Side

```
0:006> !heap -s 003b0000

Walking the heap 003b0000 ....
 0: Heap 003b0000
   Flags             00001002 - HEAP_GROWABLE
   Reserved          7424 (k)
   Commited          5720 (k)
   Virtual bytes     6240 (k)
   Free space        43 (k)
   External fragmentation           0% (3 free blocks)
   Virtual address fragmentation    8% (26 uncommited ranges)
   Virtual blocks     0
   Lock contention 15
   Segments           4
   2432 hash table for the free list
       Commits 0
       Decommitts 0
```

... Page 1/3

LFH Heap Side

```
Low fragmentation heap   003b0688
      Lock contention        4
      Metadata usage     76800
      Statistics:
          Segments created       2236
          Segments deleted        733
          Segments reused           0
          Conversions               0
          ConvertedSpace            0

      Block cache:
           Free blocks              0
           Sequence                 0
           Cache blocks      0      0     14     37     70     74     19
           Available         0      0     79    252    517    795     74


                                                             ... Page 2/3
```

Default NT Heap Side

```
0:006> !heap -s 003b0000

Walking the heap 003b0000 ....
 0: Heap 003b0000
   Flags            00001002 - HEAP_GROWABLE
   Reserved         7424 (k)
   Commited         5720 (k)
   Virtual bytes    6240 (k)
   Free space       43 (k)
   External fragmentation          0% (3 free blocks)
   Virtual address fragmentation   8% (26 uncommited ranges)
   Virtual blocks    0
   Lock contention 15
   Segments          4
   2432 hash table for the free list
       Commits 0
       Decommitts 0
```

... Page 1/3

Blocks Distribution

```
                    Default heap   Front heap
   Range (bytes)     Busy  Free    Busy   Free
-----------------------------------------------
     0 -   1024       18     83   49997   9118
  1024 -   2048      113      0       0      0
  2048 -   3072       70      1       0      0
  4096 -   5120       74      0       0      0
  8192 -   9216       19      2       0      0
 16384 -  17408        9      0       0      0
 32768 -  33792        8      0       0      0
104448 - 105472        1      0       0      0
-----------------------------------------------
  Total              312     86   49997   9118
```

Discussion