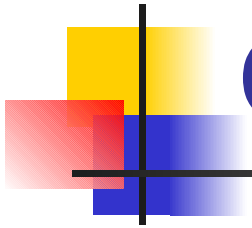
14th NASA Goddard - 23rd IEEE Conference on Mass Storage Systems and Technologies (MSST2006), College Park, Maryland USA

MRRC: An Effective Cache for Fast Memory Registration in RDMA

Li Ou
Department of Electrical and Computer Engineering
Tennessee Technological University
lou21@tntech.edu

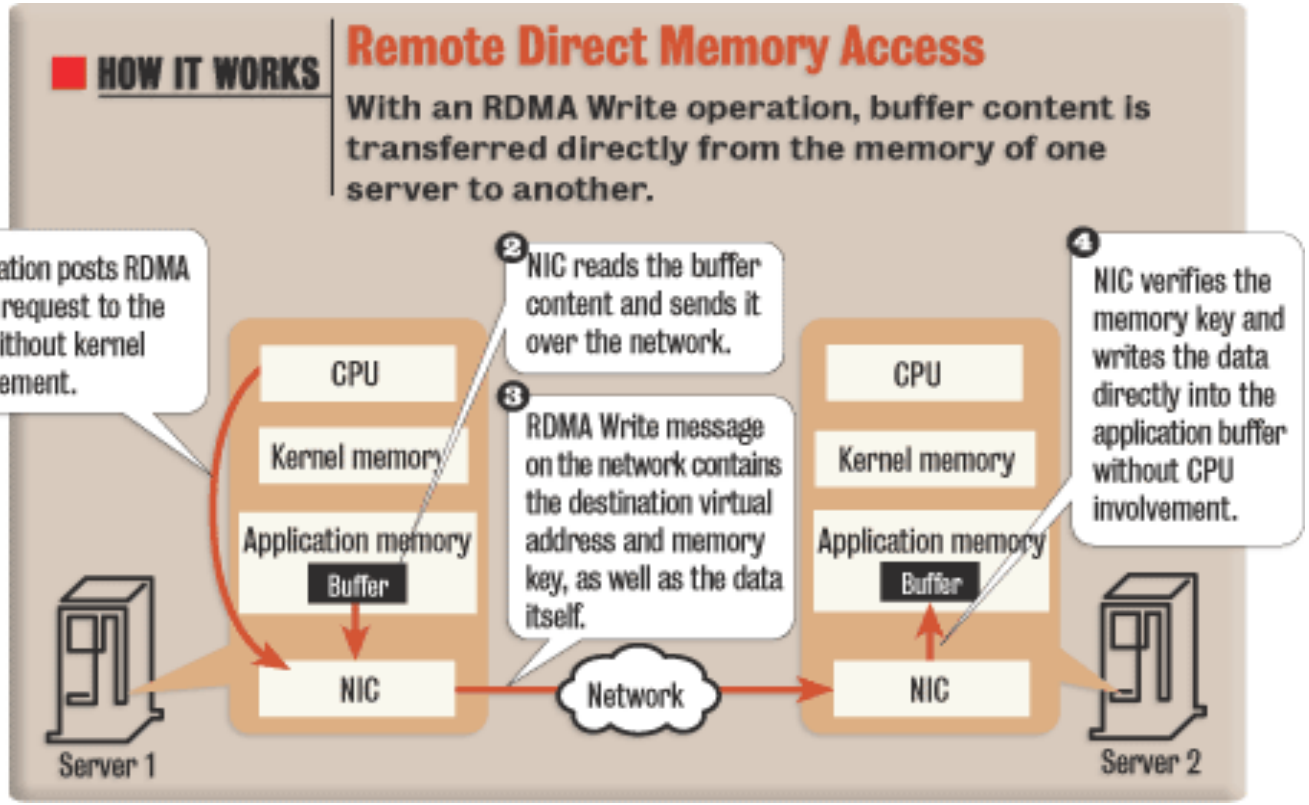
Co-authors:
Xubin (Ben) He, Tennessee Technological University
Jizhong Han, Chinese Academy of Sciences



Outline

- **Motivation**
- Design of MRRC
- Simulation results
- Conclusions

How RDMA works



RDMA offers low overhead, high speed. Ariel Cohen, Network World, 03/24/03

Memory registration of RDMA

- Why:

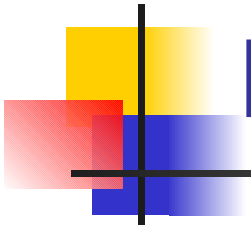
  NIC writes or reads user specified buffers directly, so before each RDMA operation, it is required to register a memory region where user buffers are located.

- How:
  - Maps the virtual memory address to the physical address.
  - Pins the memory region to make sure that it is not swapped out.
  - Writes the mapping information into a table of the NIC.
- After:
  - Deregisters the memory region.

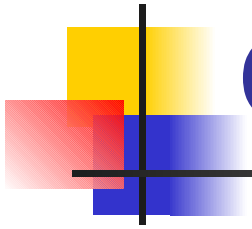
Cost of memory registration

- Memory registration and deregistration is expensive:
  - Myrinet, 4KB, Network cost is 25.6us. Registration in a Pentium Pro machine (200MHz) needs 26us.
  - Infiniband, 4KB, Network cost is 9us. Registration in a Intel Xeon 2.4GHz needs 8us.
- A model for memory registration [Wu'2003]
  - T=a*p+b
  - a: registration cost per page, b: overhead per operation, p: memory region in pages.
  - Registration: a: 0.77 us. b: 7.42us.
  - Deregistration: a: 0.22us. b: 1.1us.

MRRC: Memory Registration Region Cache



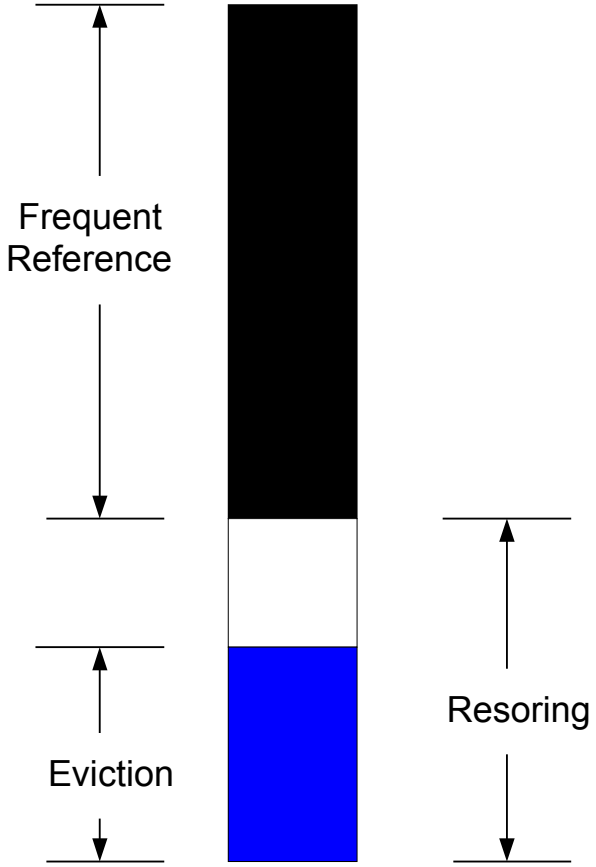
- Uses a special cache to reduce the overhead of memory registrations.
- Delays deregistration of memory regions for future possible references until they are evicted from the cache space.
- Manages cache space in terms of memory regions, instead of blocks.
- Designs a special replacement algorithm to improve hit ratios.



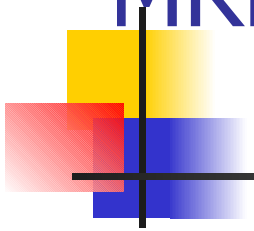
Outline

- Motivation
- **Design of MRRC**
- Simulation results
- Conclusions

LRU stack of MRRC



MRE: Memory Resorting and Eviction

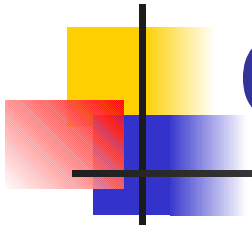


- Considers both recency and size of memory regions
- Evictfact(s,r) = r+1/s
  - s is size of a memory region.
  - r is recency.
- The smaller the Evictfact of a memory region, the closer to the bottom of the stack.

Algorithm of MRRC

```
/* procedure to be invoked upon a reference to
memory region b */

if b is in cache
   move b to the top of the stack;
else if b belongs to memory region c
    move c to the top of the stack;
else if b overlap with  memory region d {
    u = b && d;
    v = b - u;
    move d to the top of the stack;
    register v and add v to the top of the stack;
}
else
    register b and add b to the top of the stack;
```

```
r = 0;

/* procedure to be invoked upon the full of the
cache*/

/* e is the region at the bottom of the stack */

r = e.evictfact;
for each region a in resorting section
    a.evictfact = r + 1/s;    /* s is size of a region */
Resort each region in resorting section according
to evictfact;
Batched deregister all regions in Eviction
Segment;
Evict all regions in Eviction Segment;
```



Outline

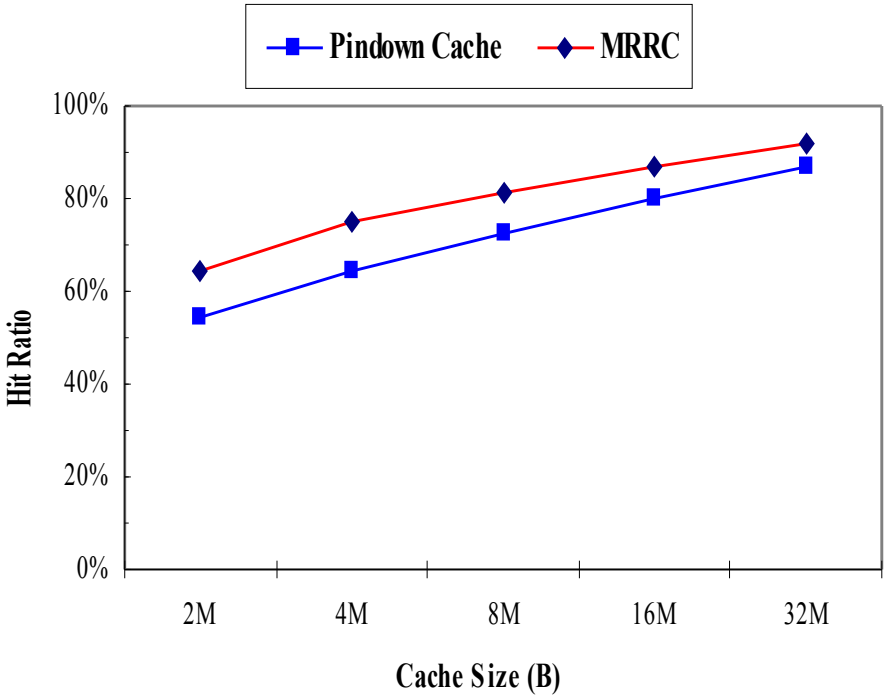
- Motivation
- Design of MRRC
- **Simulation results**
- Conclusions

Simulation methodology

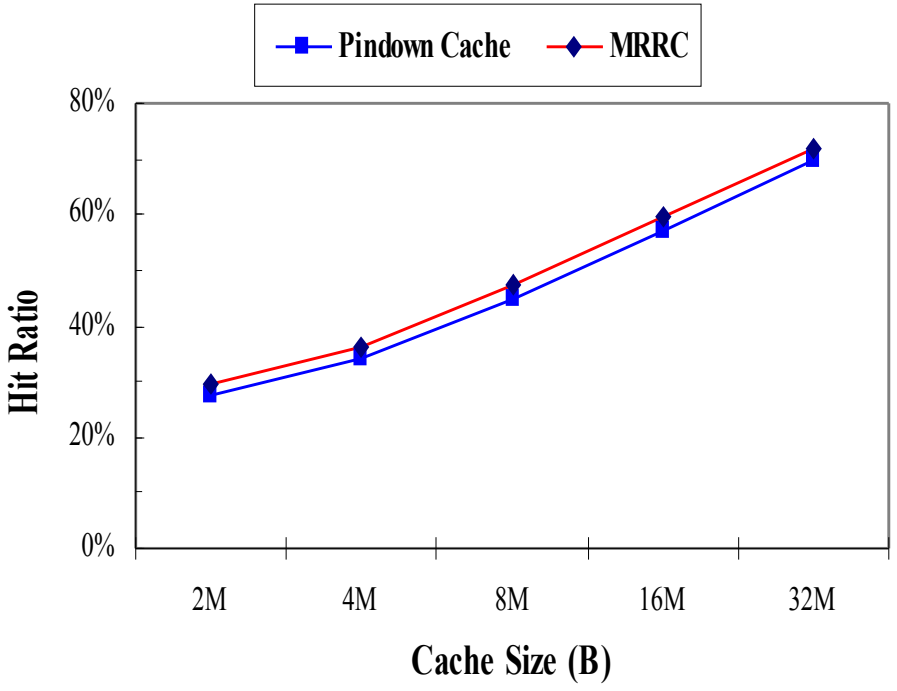
- Use trace-driven simulation to evaluate cache hit ratios.
  - Use HTTPD and Cello92 trace
- Compare hit ratios of Pindown cache and MRRC.
- Estimate response time using a mathematic model.

$$T_{total} = \sum_{i=0}^{N} ismiss(i) * T_r(i) + \sum_{j=0}^{m} T_{ur}(j)$$

Cache hit ratios

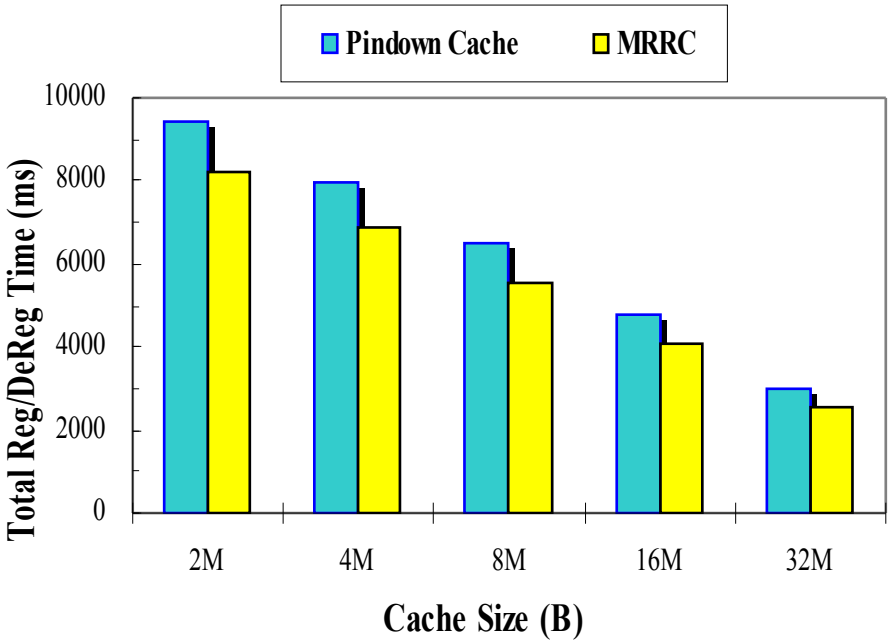


HTTPD

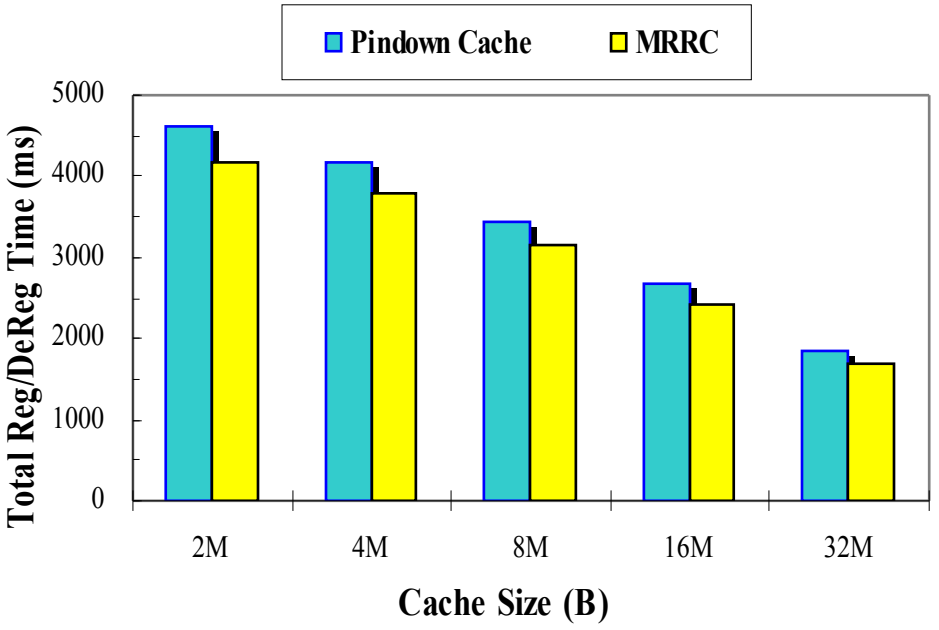


DB2

Registration cost

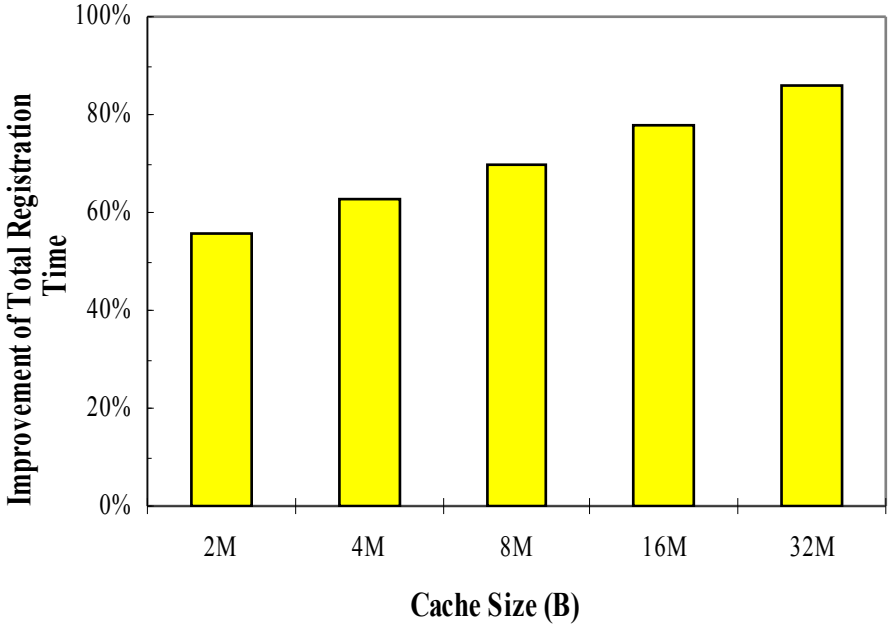


HTTPD

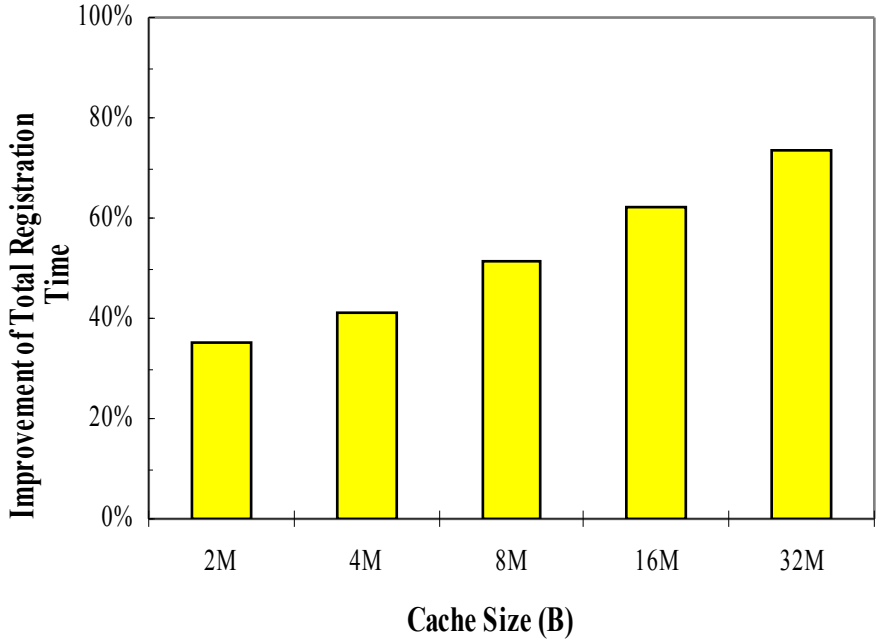


DB2

Improvement



HTTPD



DB2

Conclusions and Acknowledgments

- Conclusions:
  - MRRC reduces the overhead of memory registrations by using a special memory region cache.
  - MRE is effective because it considers both recency and size of memory regions.

- Acknowledgements:

  - Research Office and Center for Manufacturing Research, Tennessee Technological University
  - 973 Program of China
  - Faculty Research Grant at Institute of Computing Technology, Chinese Academy of Sciences

14th NASA Goddard - 23rd IEEE Conference on Mass Storage Systems and Technologies (MSST2006), College Park, Maryland USA

MRRC: An Effective Cache for Fast Memory Registration in RDMA

Li Ou
Department of Electrical and Computer Engineering
Tennessee Technological University
lou21@tntech.edu

Co-authors:
Xubin (Ben) He, Tennessee Technological University
Jizhong Han, Chinese Academy of Sciences