CCGrid2013 The 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Delft, The Netherlands

User-level Remote Memory Paging for multithreaded Applications

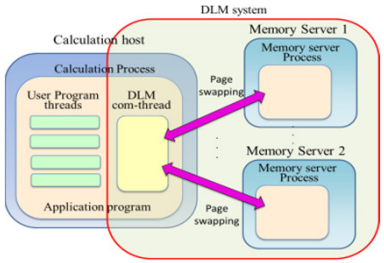
Hiroko Midorikawa, Yuichiro Suzuki, Masatoshi Iwaida, JST CREST & Seikei University, Tokyo Japan

midori@st.seikei.ac.jp http://www.ci.seikei.ac.jp/midori/paper

Background

Virtual Large Memory efficiently uses distributed memory over a cluster as a memory resource. It will be a remedy to suppress power consumption of the memory in a whole system.

Accessing remote memory from a multithreading process is a key operation not only for semi-parallel single-node programs but also for full-parallel multiple-nodes programs.



User-level remote memory paging system, ***Distributed Large Memory,*** DLM [1][2]

DLM Benefits:

- Available to solve bigger size problems using existing sequential codes without local memory size limitation
- Release users from the substantial amount of work
  - *Redesign of the sequential algorithm*
  - *Converting the existing codes to parallel MPI codes*
  - *Parallel code debugging*
- Easy to use remote memory with No knowledge of MPI
- Highly portable to various clusters, user-level software.

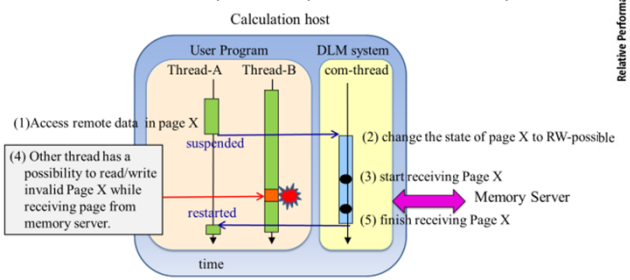
Designed for the users who prefer and accept extra execution time caused by using remote memory partially.

Remote paging for semi-Parallel and Pseudo seq. codes

to accelerate the performance of seq. codes

- Easy parallelizing by OpenMP
- Usage of implicitly multithreaded library functions

User threads are dynamically created and destroyed.



Inconsistent Page problems in user-level remote memory paging

Simple Threads Suspension : user-level implementation

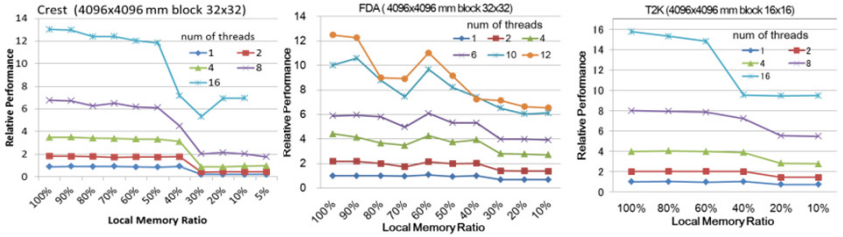
(1) The DLM hooks a pthread_creat() call and replaces it with the tailored pthread_creat() call which registers all user thread IDs.
(2) When a user thread accesses the remote data, the com-thread requests and receives a remote page to/from a memory server in background.
(3) Suspends all user thread with pthread_kill() and copy the page in a receive-buffer to user-space.
(4) Restart all user threads using pthread_kill().

Remote Memory Bandwidth (Stream benchmark, Triad)

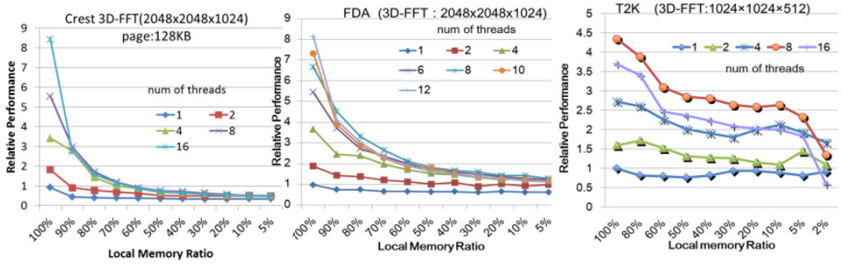
Cluster	year	NetWork	Net Spec	Local Memory (MB/s)	Remote Memery (MB/s) Ratio to LM
Crest	2013	Infiniband 4x FDR	6.8GB/s	15409	277 (2%)
FDA	2012	Infiniband 4x QDR	4GB/s	9196	399 (4%)
T2K	2009	Myri-10G x2	2.5GB/s	2700	493 (18%)
T2K	2009	Myri-10G x4	5GB/s	2700	613 (18%)
CSLM	2008	10GbEthernet/Myri10G	10Gbps	2925	380 (12%)

Why absolute remote memory bandwidth decline ?

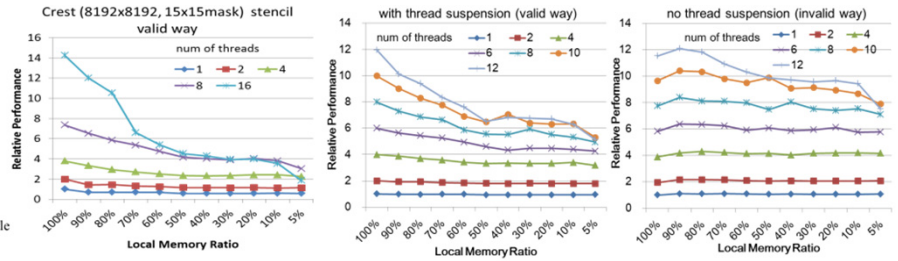
Performance : Acceptable, depends locality & cal/mRW ratio



Matrix multiplication(4095x4096)



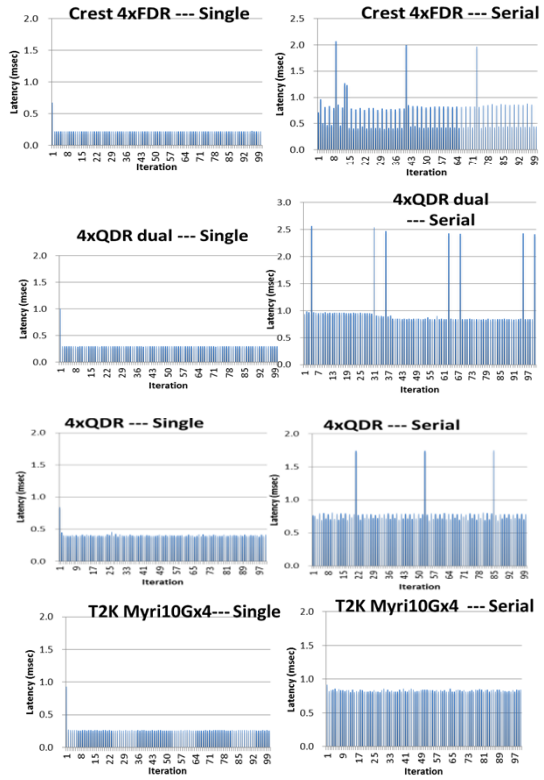
3D- FFT program using fftw multithreaded library



Thread suspension overhead in FDA cluster stencil program (8192x8192, 15x15 mask)

Real Network performance

MPI_send/Recv Ping-Pong Latency profile test



Experimental Environment

T2K cluster (16 cores/node AMD 4Core Opteron 8356 (2.3GHz) , MPICH-MX, Myrinet-10G x2)
FDA cluster (12 cores/node Xeon X5680(3.33GHz), MPICH, IB 4xQDR, IPoIB)
Crest cluster (16cores/node Xeon E5-2687W (3.1GHz) , MVAPICH, IB 4xFDR)

[1] H. Midorikawa, et.al, "Using a Cluster as a Memory Resource: A Fast and Large Virtual Memory on MPI," Proc. IEEE Cluster2009, pp.1-10
[2] H. Midorikawa, et.al, "DLM: A Distributed Large Memory System using Remote Memory Swapping over Cluster Nodes," Proc. IEEE Cluster2008, pp.268-273