vGASNet: A PGAS Communication Library Supporting Out-of-Core Processing
Ryo Matsumiya and Toshibo Endo
Tokyo Institute of Technology

Background
- Partitioned Global Address Space (PGAS) eases distributed programming.
- Out-of-core processing is required in many fields.
- Few PGAS frameworks support out-of-core processing.
- Node-local SSDs are available for out-of-core processing.

Cache replacement policy
- LRU-based policy
- Consists of pure a LRU queue and a FIFO queue.
  - The size of the FIFO queue is half of the cache pool.
  - The size of the LRU queue is the rest of the cache pool.
  1. All stored caches are enqueued to the LRU queue firstly.
  2. When the cache pool is filled, a cache is dequeued from LRU queue.
     2-a) If any other nodes have the same cache line, the cache is evicted.
     2-b) Otherwise, the cache is enqueued to the FIFO queue.
  3. When the FIFO queue is filled, the bottom of the FIFO queue is dequeued and evicted.

Performance evaluation
- Conducted with TSUBAME-KFC/BL [5].
- Page size is 4 MB.
- The cache pool size of each node is about 16 GB.

Cache mechanism overview
- SSD is much slower than DRAM.
- Under vGASNet, each node has its own cache pool on DRAM.
- Page-based cache
- Reducing the amount of accesses to the SSDs, vGASNet adopts cooperative caching mechanism.
  - Using not only local caches but also remote caches.
  - Firstly implemented in a distributed file system [2].

Cooperative caching
- Each page is originally stored in the SSD of its owner.
- Each node has a cache table, which assigns its own pages with the node whose DRAM stores the cache.
- The rough flow of forwarding cache is below.
  1. In this example, node B is to receive a page of node A.
  2. Node B requests node A.
  4. When node C has a cache of the page, node B forwards the request to node C.
  5-a) Node C sends the page cache to node B.
  5-b) When no node has any caches of the page, node A reads the original page from its SSD to the buffer.
  6-a) Node C sends the buffer to node B.
  6-b) When node B has received the cache, node B requests node A to register the cache with node A's cache table.

Guarantee cache consistency is challenging.
- In vGASNet, MOESIF protocol is implemented as a cache coherence protocol.
- MOESIF protocol is based on two practical protocol.
  1. MOESI protocol [3]
  2. Used in AMD 64-bit multi-core processors.
  3. Dirty caches are not evicted if the same page is cached in another node.
- MOSIF protocol [4]
  1. Used in Intel multi-core processors.
  2. The node whose cache is used to forward the cache is specified per cache line.

Future work
- Implementing a practical PGAS runtimes.
  - Undergoing: UPC++
- Performance evaluation using practical applications.
  - Numerical solver, machine learning, genetic analysis, etc.

Related work
- ComEx-PM [6]
  - A PGAS communication library supporting out-of-core processing.
  - The cache mechanism depends on Linux VFS cache
- HHRT [7], Papyrus [8]
  - Ease MPI programs of supporting out-of-core processing.
  - Unlike vGASNet, their interface is not based on remote memory access.

vGASNet overview
- Remote memory access based communication library.
  - The interface is similar to GASSNet [1].
- Two memory regions are available.
  - Segmented memory region
    - Can be accessed by other nodes.
    - Allocated in node-local SSDs with vgasnet_allocate()
  - Non-segmented memory region
    - Local access only.
    - Allocated in DRAM with general memory allocate functions (e.g. malloc()).

Performance evaluation

Throughput of sequential load

Throughput of random load (block size=4 MB)

Data size (MB)

Throughput (MB/s)

Future work

- Implementing a practical PGAS runtimes.
  - Undergoing: UPC++
- Performance evaluation using practical applications.
  - Numerical solver, machine learning, genetic analysis, etc.

Related work

- ComEx-PM [6]
  - A PGAS communication library supporting out-of-core processing.
  - The cache mechanism depends on Linux VFS cache
- HHRT [7], Papyrus [8]
  - Ease MPI programs of supporting out-of-core processing.
  - Unlike vGASNet, their interface is not based on remote memory access.

[1] Chae and Iqbal, Runtime Data Flow Graph Scheduling of Matrix Computations with Multiple Hardware Accelerators, IW-Novo, 50 (1996)
[5] Matsui and Endo, FPGAS Communication Runtime for Extreme Large Data Computations, in Proc. of ESREF2 '16
[6] Endo, Realizing Out-of-Care Stencil Computations using Multi-Thread Memory Hierarchy on GPU/CPU Clusters, in Proc. of Cluster '16

This work is supported by 1ST-CREST, "Software Technology that Deals with Deeper Memory Hierarchy in Post-petascale era".