Implementing Global Memory Management in a Workstation Cluster (GMS) - CS739 Class Discussion 2/26/2004

Todd Jones
stjones@cs.wisc.edu
March 8, 2004

1 GMS Review

1.1 Goal
Globally coordinate memory management among mutually trusting nodes in a networked cluster.

1.2 Algorithm

• Two types of pages exist.
  – Global pages are stored in the virtual memory of a cluster node on behalf of the entire cluster
  – Local pages are pages that have been recently used by the local node.
    * Shared pages are local to more than one node.

• When a node (A) faults on a non-resident page it fetches the page from the local or global memory of another cluster node into a local page of its own if possible.

• If the page is not available in any of the cluster member’s memories, the page will be read from disk as usual.

• In order to maintain balance among systems, the faulting node (A) will lazily evict pages from its own memory (local or global) to global memory on a system with idle memory (i.e. on a system with old global pages).
• The previous step requires nodes to have some global notion of the age of pages in the cluster.
  – An epoch based, centrally coordinated protocol allows a node assigned to be the coordinator for a specific epoch to collect idleness information from each cluster node and redistribute it throughout the cluster.
  – Cluster nodes select target nodes upon which to place evicted pages using a weighted probabilistic algorithm. The weights used in the algorithm are provided by the globally distributed page age information. I.e. a cluster node with more idle pages will have a relatively larger weight and so will receive proportionally more requests to store pages by nodes evicting pages from their own memories.

1.3 Data Structures

Various local and distributed data structures are described that GMS uses to locate the cluster node that is currently caching a memory page.

• The page ownership directory (POD) is a globally replicated data structure that maps page unique ids to the GCD section containing cache information about the page.

• The global cache directory (GCD) stores network location information about cached pages in the cluster. It is a distributed hash table. Each node in the cluster maintains a portion of the table.

• The page frame directory is a local data structure storing physical page location and age information for pages in a node’s local and global memories.

1.4 Page Lookup

Figure four in the paper demonstrates the three step page lookup process for locating a cached page in the cluster.

• Given a page’s unique id, consult the POD to determine what section of the GCD corresponds to it.

• Consult the GCD entry (potentially over the network) to determine if the page is cached and where.

• The GCD node contacts the node currently caching the page (potentially over the network) and requests that it send the page to the faulting node. The GCD is updated with the new location of the page in the cluster (i.e. the faulting node).
2 Class Discussion

2.1 Differences Between GMS and Cooperative Caching

- GMS takes more of the distributed systems issues into account (like failure and load variability). Since it is a real implementation, it makes sense that this is true.
- GMS uses a weighted probabilistic algorithm for target selection vs the uniformly random distribution used by N-chance forwarding.
- GMS centrally collects and re-distributes idleness information to generate weights for the target host selection algorithm.
- Epochs are used to limit the lifetime of idleness information.
- Tries to be smarter about allocation decisions.
- Different evaluation style.
  - Application benchmarks vs. aggregate workload traces.
    * There is no sense for how the applications included in the benchmarks are actually used and no justification is made for their particular use in the evaluation. For example, is the typical use very interactive and bursty or uniform?
    * Shows that in the case of non-uniformly distributed and scarce idleness in a cluster, GMS outperforms N-chance forwarding substantially in finding idle pages.
    * No justification is given for highlighting this condition. Is idle memory really that hard to find in a cluster like the one they are targeting?
    * Does not address scalability. It may be the case that N-chance forwarding will out-perform GMS due to its lack of centralized coordination.

3 Summary

Coordinated caching and GMS can make something that is slow (paging due to memory shortage) slightly less slow by utilizing remote memory. If you have a workload that pages, wouldn’t you just buy more memory so that it doesn’t page? These systems lack a strong motivational power even though they are technically interesting.