Cluster Computing

Network RAM

- Introduction
- Remote Memory Paging
- Network Memory File Systems
- Applications of Network RAM in Database
- Summary

Introduction

- Discuss efficient distribution & exploitation of main memory in a cluster
- Network RAM or Network Memory
  - the aggregate main memory of the workstation in a cluster
  - local memory vs. remote memory
- A significant amount of resources unused in the cluster at any given time
  - 80-90% of workstation will be idle in the evening & late afternoon
  - 1/3 of workstations are completely unused, even at the busiest times of day
  - 70-85% of the Network RAM in a cluster is idle
- Consist of volatile memory
  - but still can offer good reliability
    - use some form of redundancy, such as replication of data or parity
  - offer an excellent cost-effective alternative to magnetic disk storage

Introduction

- Technology trend
  - High performance networks
    - switched high performance local area network : ex) myrinet
    - low latency communication and messaging protocols : ex) U-Net
      - latency of Active Message on ATM (20 μs) vs. magnetic disk latency (10ms)
  - High performance workstations
    - memories with low cost
      - Network RAM even more attractive
  - The I/O processor performance gap
    - the performance improvement rate for
      - microprocessor: 50-100% per year
      - disk latency: 10% per year
      - network latency: 20% per year
      - network bandwidth: 45% per year

Memory Hierarchy Figures for a Typical Workstation Cluster

Memory Hierarchy	Latency (μsec)	Bandwidth (MB/s)	Capacity (MB)
Cache	0.002	500	2
DRAM	0.1	200	256
Network RAM	20	15	12800
Disk	10000	10	8000

- A typical NOW with 100 workstations
- A network of 20 μs latency & 15 MB/s bandwidth

Introduction

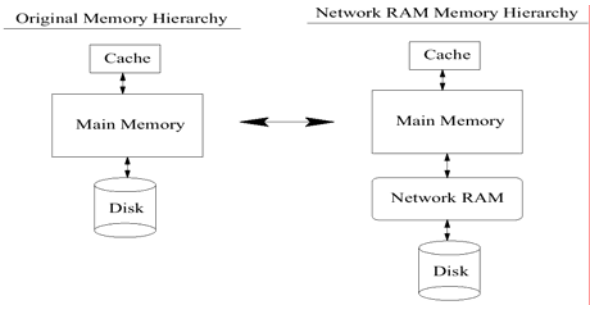
- Issues in Using Network RAM
  - Using network RAM consists essentially of locating & managing effectively unused memory resources in a workstation cluster
    - keeps up-to-date information about unused memory
    - exploit memory in a transparent way so that local users will not notice any performance degradation
    - require the cooperation of several different systems
  - Common use of Network RAM
    - Remote memory paging
      - performance: between the local RAM and the disk
    - Network memory filesystems
      - can be used to store temporary data
        - by providing Network RAM with a filesystem abstraction
    - Network memory DBs
      - can be used as a large DB cache and/or a fast non-volatile data buffer to store DB sensitive data

Remote Memory Paging

- 64-bit address space
  - heavy memory usage
  - application's working sets have increased dramatically
    - sophisticated GUI, multimedia, AI, VLSI design tools, DB and transaction processing systems
- Network RAM
  - equivalent of remote memory paging
  - (cache) – (main memory) – (network RAM) – (disk)
  - faster than the disk
  - all memory-intensive applications could benefit from Network RAM
  - useful for mobile or portable computers, where storage space is limited

7

Changing the Memory Hierarchy



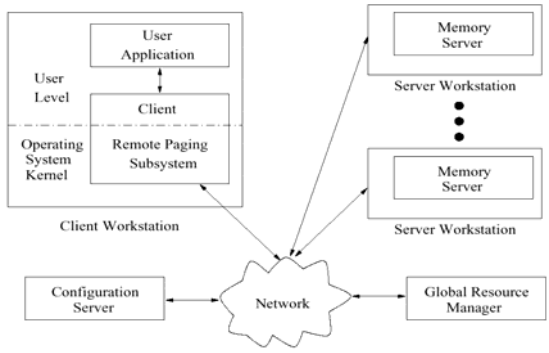
8

Remote Memory Paging

- Implementation Alternatives
  - Main idea of remote memory paging
    - start memory server processes on workstations that are either idle or lightly loaded and have sufficient amount of unused physical memory
  - A policy for the page management
    - which local pages?
    - which remote nodes?
      - client periodically asks the global resource manager
      - distribute the pages equally among the servers
    - how server load?
      - negotiation from server to client & migration
      - transfer data to server's disk

9

Remote Paging System Structure



10

Remote Memory Paging

- Implementation Alternatives
  - the client workstation has a remote paging subsystem
    - user-level
    - in the OS
      - enable it to use network memory as backing store
  - a global resource management process
    - running on a workstation in the cluster
    - hold information about the memory resources & how they are being utilized in the cluster
  - a configuration server (registry server) process
    - responsible for the setup of the network memory cluster
    - is contacted when a machine wants to enter or leave the cluster
    - authenticated according to the system administrator's rules

11

Remote Memory Paging

- Client remote paging subsystem
  - 2 main components
    - a mechanism for intercepting & handling page faults
    - a page management policy
      - keep track of which local pages are stored on which remote nodes
  - 2 alternative implementations of transparent remote paging subsystem on the client workstation
    - using the mechanisms provided by the existing client's OS
    - by modifying the virtual memory manager & its mechanisms in the client's OS kernel

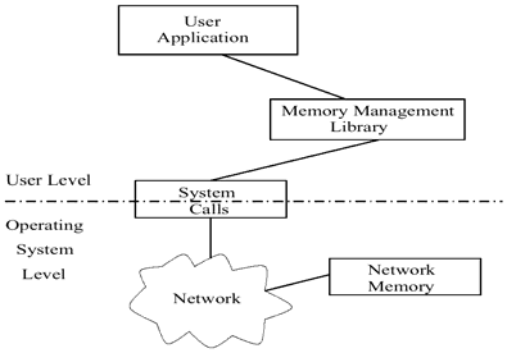
12

Transparent Network RAM Implementation using Existing OS Mechanisms

- User Level Memory Management
  - each program uses new memory allocation & management calls, which are usually contained in a library dynamically or statically linked to the user applications
  - require some program code modifications
  - prototype by E. Anderson
    - a custom *malloc* library
    - using TCP/IP, perform a 4K page replacement in 1.5 – 8.0 ms (1.3 – 6.6 faster)
  - mechanism
    - intercept page faults through the use of segmentation faults
    - a segmentation fault signal from OS memory manager
    - this signal intercepted by the user level signal handler routines
    - page management algorithm is called to locate & transfer a remote page into local memory, replacing a local page

13

User Level Memory Management Solution's Structure



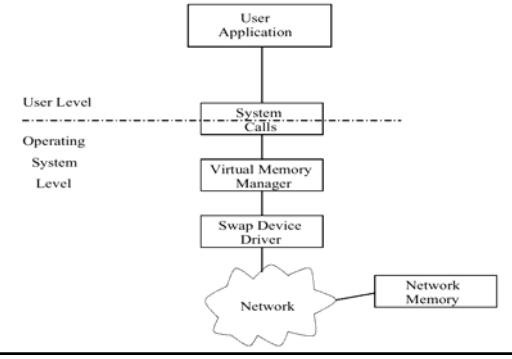
14

Transparent Network RAM Implementation using Existing OS Mechanisms

- Swap Device Driver
  - the OS level implementation
  - all modern operating system provide virtual memory manager which swap pages to the first device on the list of swap devices which will be the remote memory paging device drivers
  - create a custom block device & configure the virtual memory manager to use it for physical memory backing store
    - when run out of physical memory, it will swap pages to the first device on the list (may be the remote memory paging device driver)
    - if and when this Network RAM device runs out of space, the virtual memory manager will swap pages to the next device
  - simple in concept & require one minor modification to the OS
    - the addition of a device driver

15

Swap Device Driver Solution's Structure



16

Transparent Network RAM Implementation using Existing OS Mechanisms

- Page Management Policies
  - which memory server should the client select each time to send pages
    - global resource management process
    - distribute the pages equally among the servers
  - handling of a server's pages when its workstation becomes loaded
    - inform the client & the other servers of its loaded state
    - transfer them to the server's disk & deallocate them each time they are referenced
    - transfer all the pages back to the client
  - when all servers are full
    - store the pages on disk

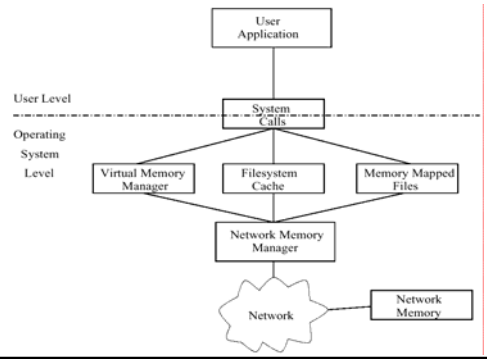
17

Transparent Network RAM Implementation Through OS Kernel Modification

- OS modification
  - software solution offers the highest performance without modifying user program
  - an arduous task
  - not portable between architectures
- Kernel modification
  - provide global memory management in a workstation cluster
  - using a single unified memory manager as low-level component of OS that runs on each workstation
  - can integrate all the Network RAM for use by all higher-level functions, including virtual memory paging, memory mapped files and filesystem caching

18

Operating System Modification Architecture



19

Transparent Network RAM Implementation Through OS Kernel Modification

- Global memory management
  - Page fault in node P in GMS
    - the faulted page is in the global memory of another node Q
    - the faulted page is in the global memory nodes of node Q, but P's memory contains only local pages
    - the page is on disk
    - the faulted page is a shared page in the local memory of another node Q

20

Boosting Performance with Subpages

- Subpages are transferred over the network much faster than whole pages
  - not with magnetic disk transfers
- Using subpages allow a large window for the overlap of computation with communication

21

Reliability

- Main disadvantages of remote memory paging
  - security: can be resolved through the use of a registry
  - reliability (fault-tolerant)
- Use a reliable device
  - the simplest reliability policy
  - reliable device such as the disk
    - async write but, bandwidth limited
    - can achieve high availability but simple
    - performance under failure
    - disk memory overhead (expensive)
- Replication – mirroring
  - using page replication to remote memory without using disk
  - (-) use a lot of memory space, high network overhead

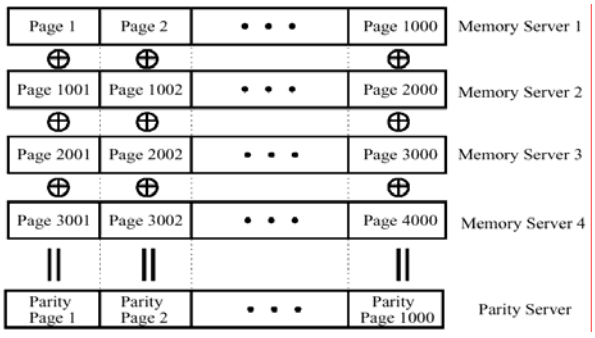
22

Reliability

- Simple parity
  - extensively used in RAIDs, based on XOR operation
  - reduce memory requirement (1+1/S)
  - network bandwidth & computation overhead
  - if two servers have failure???
- Parity logging
  - reduce the network overhead close to (1 + 1/S)
  - difficult problems from page overwriting

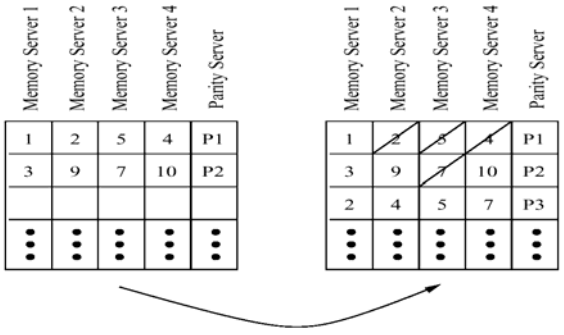
23

An Example of the Simple Parity Policy (Each server here can store 1000 pages)



24

A Parity Logging Example
with Four Memory Servers and One Parity Server



Remote Paging Prototypes

- The Global Memory Service (GMS)
  - A modified kernel solution for using Network RAM as a paging device
  - (reliability) Write data asynchronously to disk
  - The prototype was implemented on the DEC OSF/1 OS
    - under real application workloads, show speedups of 1.5 to 3.5
    - respond effectively to load changes
  - Generally very efficient system for remote memory paging

Pros and Cons of GMS

+	+ Comprehensive exploitation of Network RAM in every level of the operating system (virtual memory paging, memory mapped files and filesystem buffering) + Global memory management in the NOW + Effective memory load balancing + Good performance
-	- Major changes to the operating system - Requires a homogeneous workstation cluster - Uses the disk for reliability

Remote Paging Prototypes

- The Remote Memory Pager
  - use the swap device driver approach
  - the prototype was implemented on the DEC OSF/1 v3.2 OS
  - consist of a swap device driver on the client machine & a user-level memory server program runs on the servers
  - effective remote paging system
    - no need to modify the OS

Pros and Cons of the Remote Memory Pager

+	+ Good exploitation of Network RAM for paging + Does not modify the OS, thus it can be implemented in any commercial operating system + Good performance + Operates in a heterogeneous workstation cluster + Implements effective reliability policies without the disk
-	- No global memory management in the NOW - Uses Network RAM only for paging and not as a file cache or for memory mapped files.

Network Memory File System

- Use the Network RAM as a filesystem cache, or directly as a faster-than-disk storage device for file I/O
- Using Network Memory as a File Cache
  - all filesystem use a portion of the workstation's memory as a filesystem cache
  - two problem
    - multiple cached copies waste local memory
    - no knowledge of the file's existence
  - several ways to improve the filesystem's caching performance
    - eliminate the multiple cached copies
    - create a global network memory filesystem cache
      - 'cooperate caching'

Network Memory File System

- Network RamDisks
  - Disk's I/O performance problem
  - Use reliable network memory for directly storing temporary files
  - Similar to Remote Memory Paging
  - Can be easily implemented without modifying OS (Device Driver)
  - Can be accessed by any filesystem (NFS)
  - Network RamDisks
    - a block device that unifies all the idle main memories in a NOW under a disk interface
    - behaves like any normal disk, but implemented in main memory (RAM)
    - diff
      - instead of memory pages, send disk blocks to remote memory
      - disk blocks are much smaller than memory pages

31

Applications of Network RAM in Databases

- Transaction-Based Systems
  - to substitute the disk accesses with Network RAM accesses
    - reducing the latency of a transaction
  - a transaction during its execution makes a number of disk accesses to read its data, makes some designated calculations on that data, writes its results to the disk and, at the end, commits
    - atomicity & recoverability
  - 2 main areas where Network RAM can be used to boost a transactions-based system's performance
    - at the initial phase of a transaction when read requests from the disk are performed
      - through the use of global filesystem cache
    - speed up synchronous write operations to reliable storage at transaction commit time
      - transaction-based systems make many small synchronous writes to stable storage

32

Applications of Network RAM in Databases

- Transaction-Based Systems
  - Steps performed in a transaction-based system that uses Network RAM
    - at transaction commit time, the data are synchronously written to remote main memory
    - concurrently, the same data are asynchronously sent to the disk
    - the data have been safely written to the disk
  - In the modified transaction-based system
    - a transaction commits after the second step
    - replicate the data in the main memories of 2 workstations
  - Results
    - the use of Network RAM can deliver up to 2 orders of magnitude higher performance

33

Summary

- Emergence of high speed interconnection network added a new layer in memory hierarchy (Network RAM)
- Boost the performance of applications
  - remote memory paging
  - file system & database
- Conclusion
  - Using Network RAM results in significant performance improvement
  - Integrating Network RAM in existing systems is easy
    - device driver, loadable filesystem, user-level code
  - The benefits of Network RAM will probably increase with time
    - gap between memory and disk continue to be widen
- Future Trend
  - Reliability & Filesystem Interface

34