Scalable Supercomputing Comparing MPPs and SuperClusters™

R. Kent Koeninger, Cray Research, Inc., 655F Lone Oak Drive,
Eagan, Minnesota 55121

ABSTRACT: *This paper compares the characteristics of loosely coupled clusters and tightly coupled MPPs.*

1 Introduction

In this paper I review ideas for scalable computing and show how these apply to Cray Research MPPs and SuperClusters™. I contrast characteristics of capability and capacity systems and indicate how these techniques will be merged in future scalable architectures.

2 Concepts

"Scalable computing" allows one to upgrade computational power over a large range of performance, while retaining compatibility. The scalability depends on the distance (time) between nodes in a parallel system.

Latency is the time required to send one byte between nodes. Tightly coupled systems have short latencies (a few microseconds) and are often used for capability solutions. Loosely coupled systems have long latencies (a few milliseconds or many microseconds) and are often used for capacity solutions.

A capability system allows one to run a single very large job, sustaining many computations per second, or using many gigabytes of memory. Short communication latencies are generally required to sustain a high performance across distributed-parallel nodes. Applications that can exploit a loosely coupled system in a capability mode either communicate rarely among nodes or use asynchronous communication, overlapping the long communication times with long computations. These codes are sometimes called "embarrassingly parallel."

A capacity system provides high throughput for many jobs running at once. Performance is generally measured in computations per hour across many jobs. Loosely coupled systems with long communication latencies work well in this environment, when many jobs run in parallel, each on one node at a time. The capability is limited by the size of the largest node and the capacity is the sum of the capabilities.

3 Clusters and MPPs

Clusters tend to be collections of workstations on loosely coupled networks. The trend in clusters is to increase the processing power per node by using symmetric multiprocessors and to decrease the latency between nodes. Individual jobs tend to run well on individual nodes.

MPPs tend to have many small nodes and excellent communication among nodes. Each node tends to be too slow for large capabilities, but the fast communication allows many nodes to be used in parallel for a single large application.

4 Cray SuperClusters™

Cray's current offering of SuperClusters provides a low entry cost with scalability to large capabilities and large capacities. A single node in the SuperCluster can range from a CRAY J916/4 system with less than a GFLOP/s of performance to a CRAY T932 system with more than 50 GFLOP/s of performance. The cluster can scale from as single node to as many nodes as are feasible in one's budget. Most sites tend to cluster two to eight Cray systems in one SuperCluster.

The nodes are connected with HIPPI channels, which sustain near 100 MB/s. NQE provides excellent load leveling. By combining the file caching of the Distributed File System (DFS) with the high-bandwidth Shared File System (SFS), Cray SuperClusters provide high-performance access to physically shared data.

SFS transfers data at over 80 MB/s, which is an order of magnitude faster than shared data rates on non-Cray clusters. SFS also provides high resiliency; any node in the cluster can fail while all data in on the shared disks remains available to the remaining nodes.

5 Single Node Clusters

Each node in a Cray SuperCluster tends to be faster than the capacity of most clusters of workstations in the field today. CRAY J90 systems will support up to 32 processors in one symmetric-memory node (6 GFLOP/s). Thus, a single CRAY

J92 will provide a higher capability and greater capacity than most clusters in the field today. The communication among processors is very-tightly coupled (a few microseconds) through shared memory.

Cray Research will support "shared memory" message-passing libraries to allow these symmetric-memory nodes to be programmed logically as distributed-clustered nodes. This combines the portability of message passing with the performance of shared memory.

6 Cray MPPs

Cray MPPs provide a highly scalable capability for a wide range of applications. This is a good solution when capabilities of 10s or 100s of GFLOP/s are required. The low-latency, high-bandwidth internode communication is the best in the industry, allowing many parallel applications to scale to hundreds of processors without suffering large communication overheads. These MPPs provide an excellent path for highly parallel solutions on the current generation of MPPs and on future generations of Cray parallel machines.

7 Parallel Programming

Cray SuperClusters support 32 processors on a single shared memory that can be programmed using AutoTasking$^{\mathrm{TM}}$ or message passing. These programming methods are efficient within individual SuperCluster nodes. Among nodes, one uses message passing, and the application must be tuned to account for longer communication latencies across the HIPPI channels.

CRAY MPPs are programmed using explicit communication through message passing or implicit communication through the CRAFT programming model. The explicit communication provides greater control, generally resulting in higher performance. The resulting code tends to be more portable than when implicit communication is used. On the other hand, implicit communication (CRAFT) simplifies the programming task, reducing the time needed to port codes to MPPs.

Message passing is the only programing method common to SuperClusters and MPPs. Cray Research currently supports PVM message passing and plans to support MPI message passing.

8 Futures

In each succeeding generation, one can expect Cray Research to increase the node speeds and to improve the communication speeds among nodes. This will be the case for Cray SuperClusters and CRAY T3E systems (and their successors). Cray will replace the HIPPI communication in its SuperClusters with the new SCX I/O and will boost the performance of the CRAY J90 nodes. Future generations of Cray MPPs (E.g., the CRAY T3E systems) will also have faster nodes, faster interconnects, and will use SCX for I/O.

In the long-term, Cray plans to combine these architectures using the tightly coupled distributed-memory communications of the MPP systems and the high-performance symmetric-memory parallelism of the individual SuperCluster nodes. This future architecture is code named "Scalable Node."

9 Summary

Distributed parallel systems can be used as capability or capacity engines. Cray Research offers systems to meet these varying needs. SuperClusters provide 32 processor capabilities (6 GFLOP/s per CRAY J932 system) that can be clustered into hundreds of processors (10s of GFLOP/s) of capacity. MPPs offer many hundreds of processors of capability (10s or 100s of GFLOP/s). Cray will improve the communications and node performance of each succeeding generation of SuperClusters and MPPs, eventually combining these technologies into Scalable Nodes.