2006 R&D 100 Award Winning Compute Process Allocator (CPA)

Vitus Leung, Michael Bender (Stony Brook), David Bunde (UIUC/Knox), Kevin Pedretti, Cynthia Phillips

The CPA's principal application is to maximize throughput on massively parallel supercomputers by managing how processors are assigned to particular computing jobs, given a stream of computing tasks submitted to a job queue. The CPA assigns each job to a set of processors, which are exclusively dedicated to the job until completion. The CPA obtains maximum throughput by choosing processors for a job that are physically near each other, minimizing communication and bandwidth inefficiencies. See Figure 1.

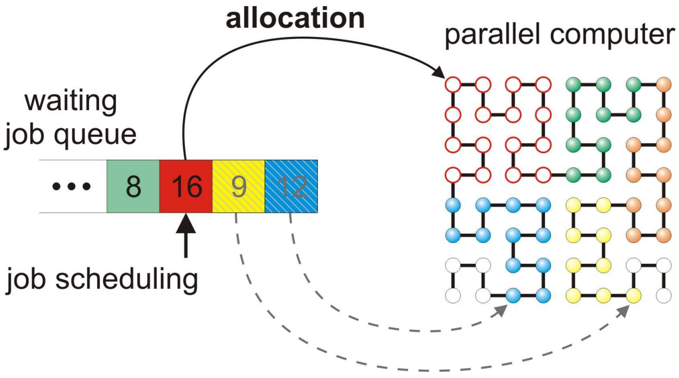

Figure 1: Allocation

In experiments at Sandia National Laboratories, the optimized node allocation strategy employed by CPA increased throughput by 23%, in effect processing five jobs in the time it normally took to process four.

The CPA is the only allocator decoupled from the compute nodes and distributed from the main resource manager. The CPA is scalable to over 10,000 nodes. Allocators coupled to the compute nodes and not distributed from the main resource manager are scalable to only 4,096 nodes.

Remarkably, the CPA can achieve processor locality for a stream of jobs in massively parallel supercomputers using simple, one-dimensional allocation strategies. The CPA accomplishes this reduction using a space-filling curve, which imposes an ordering on the network of processors such that locations near each other on the curve are also near each other in the physical network of processors. The CPA's approach is applicable even when the processors are connected by irregular, higher dimensional networks. See Figure 2.

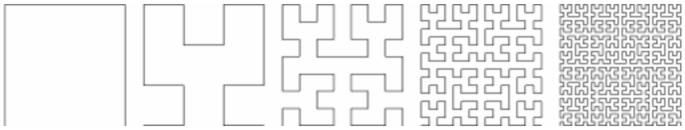

Figure 2: Hilbert Space-Filling Curve

Although simulations showed that a multidimensional strategy could give a better allocation for a single job, simulations also showed that the one-dimensional strategies employed by CPA performed better on a stream of jobs. The locally better decisions of these higher-dimensional allocation strategies seemed to paint the algorithm into a corner over time.

After extensive prototype development, the CPA was licensed to Cray Inc. in June 2005. The CPA can be ported to any system that uses a variant of NASA's Portable Batch System (PBS), and PBS Pro maintains code to interface with the CPA. PBS is the defacto standard in cluster and parallel system resource management. In addition to Cray, PBS is used to manage systems built by Dell, Hewlett Packard, IBM, and Silicon Graphics. PBS is supported on most Linux, UNIX, Windows, and Mac OS X based systems.

The development cost of the CPA was less than 1% of the development cost of a parallel computer. CPA is an example of how a small investment in computer algorithms can dramatically leverage the return on a large investment in computer hardware.

Contact:
Vitus Leung, 505-844-1896,
vjleung@sandia.gov.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.