CEARCH: Cognition Enabled Architecture

Stephen P. Crago, USC/ISI, crago@isi.edu
Janice Onanian McMahon, USC/ISI, jmcmahon@isi.edu
Chris Archer (Northrop Grumman), Krste Asanovic (MIT), Richard Chaung (US Army I2WD), Keith Goolsbey (Cycorp), Mary Hall (USC/ISI), Christos Kozyrakis (Stanford University), Kunle Olukotun (Stanford University), Una-May O'Reilly (MIT), Rick Pancoast (Lockheed Martin), Viktor Prasanna (USC), Rodric Rabbah (MIT), Steve Ward (MIT), Donald Yeung (University of Maryland)

Introduction

The goal of the CEARCH project is to develop a new computer architecture framework that will enable the deployment of cognitive systems into DoD platforms. In the past, cognitive systems have been limited because many reasoning and learning techniques could not keep up with real-time data, especially in embedded systems. Cognitive algorithms have traditionally been developed on commodity computer architectures, which have been based on sequential instruction streams with coarse-grain parallelism available on cluster-based multiprocessors. The CEARCH architecture will not be constrained by legacy sequential instruction streams and will leverage recent research on tile-based architectures, intelligent memory systems, interconnect networks, run-time systems, languages, and compilers. Our approach is to develop an introspective computer architecture driven by probabilistic and symbolic reasoning and learning techniques. The computer architecture is based on the stored processor concept, support for soft computing, and an adaptive memory system that is based on transaction-based coherence and consistency. We expect the technology developed by CEARCH to benefit a wide range of DoD missions, but CEARCH will focus on three applications domains: UAV mission planning, sea-based ballistic missile defense, and unattended ground sensors.

Algorithm and Application Characteristics

The CEARCH architecture is being driven by anticipated application and algorithm characteristics and requirements. The goal of the CEARCH architecture is to support a wide range of cognitive systems, not to dictate a particular cognitive processing approach. The CEARCH team has chosen a range of algorithms that we believe are likely to be important parts of a cognitive system. The algorithms include symbolic reasoning and learning algorithms, probabilistic reasoning and learning algorithms, and sub-symbolic reasoning and learning algorithms. Architecture requirements are not determined only by algorithms; they are also determined by data sets and the interaction of algorithms within a larger system. Table 1 summarizes our study of algorithm and system characteristics.

Table 1: Algorithm Characteristics.

Kernel	**Example Scenario System Requirement**	**Architectural drivers**
Probabilistic Relational Model (Learn, Infer)	1-2 Tera-updates/sec on large graphs	Probabilistic computation, large densely connected graphs, indirection over graph nodes, varying granularity, load balancing, trade-off: error for latency
SATisfiability-based Planner	1 Giga-Boolean-inferences per second	Parallel tree traversal, symbolic matching, partial results sharing and communication, any-time solutions, load balancing, trade-off: latency vs. search
Support Vector Machine Classification	2 Tera-ops (variable-precision floating point) / sec on sparse vectors	Variable-precision arithmetic on sparse vectors, flexible caching, computational density, trade-off: accuracy vs. support vector computation
Information-form Data Association Tracking	2 Tera-ops (probability calculations) / sec on sparse matrix	Probabilistic computation, parallel sparse matrix calculations, load balancing, trade-off: update frequency vs. accuracy
Symbolic Reasoning and Learning	313K problem trees per second	Dynamic parallel tree searches, symbolic matching, irregular memory accesses, any-time solutions, load balancing, trade-off: search time vs. completeness
System		**Rapid High-Level Reorganization and Responsivity**

Architecture

The driving principal of the CEARCH architecture is that it must support introspection and self-management. Cognitive systems are inherently dynamic; they must be able to react robustly to changing situations. Therefore, a computer architecture that supports cognitive processing must be able to perform self-management, so that it can allocate and utilize computing resources effectively in support of cognitive processing. Effective self-management requires introspection; the computer architecture must be able to

Effort sponsored by the Defense Advanced Research Project Agency (DARPA) through the Department of Interior National Business Center under grant number NBCH104009.

monitor its own performance, at all levels of the architecture. Furthermore, this introspection must be support the communication of requirements and performance data between the application programmer, run-time system, compilers, and hardware.

In order to achieve the performance required to support real-time cognitive processing, the CEARCH architecture must support parallelism. Traditional signal processing and high-performance computer architectures also support parallelism, but the parallelism in a computer architecture for cognitive processing must support parallelism at varying granularities and must support exceptionally fast context-switching in support of load balancing and resource reallocation. The processing elements themselves must be optimized for new data types and varying precision. The stored processor concept allows all processor state to migrate throughout the memory hierarchy and provides fine-grained protection between threads, and is the basis for the CEARCH processing nodes. The CEARCH architecture is envisioned to have hundreds of processing elements per chip and millions of virtual cognitive processing elements.

Cognitive systems are generally used to solve problems that are too difficult to be solved optimally or exactly, so they must be designed to tolerate approximations and non-zero error rates. A hardware computer architecture can be designed to leverage these approximations, which we call soft computing. The CEARCH architecture is being designed to take advantage of characteristics of soft computing at all levels: in the processing elements, the memory systems, communication protocols, and in the system software.

The CEARCH architecture has an adaptive memory system that is being designed to handle the challenging memory access characteristics of cognitive workloads. Cognitive processing requires irregular accesses to large working sets and also requires dynamic allocation of memory elements between processor elements and inexpensive rollback for error tolerance and conflict resolution. The CEARCH adaptive memory architecture is based on transactional coherence and consistency and also incorporates Mondriaan memory techniques and cell isolation and protection. Figure 1 summarizes the characteristics of the CEARCH hardware architecture.

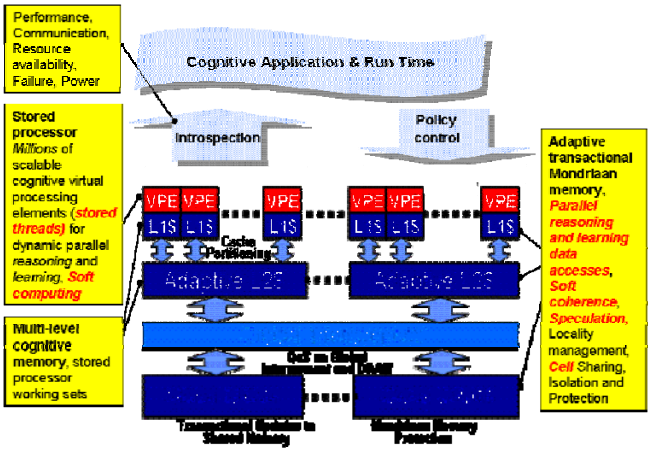

Figure 1: Summary of CEARCH Hardware Architecture

We have developed a simulator for the CEARCH architecture and have characterized the performance of several benchmarks. Figure 2 shows the performance of the CEARCH architecture on loopy believe propagation. The different bars for each number of CPUs shows the performance for the architecture taking advantage of different levels of soft computing. We have obtained speedups of up to 333x over single CPU systems without support for soft computing.

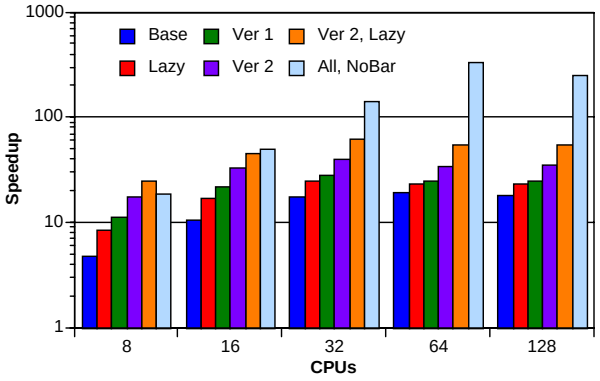

Figure 2: CEARCH Performance on Loopy Belief Propagation

System Software

The CEARCH team is also developing system software, which includes a programming model and runtime system. The system software is summarized in Figure 3.

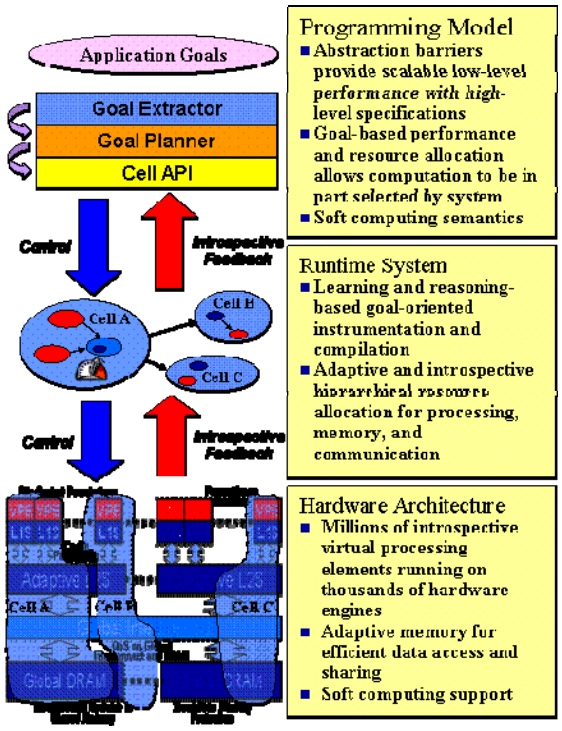

Figure 3: Summary of CEARCH System Software

Summary

The CEARCH team is developing a computer architecture, which includes hardware and system software, that will support real-time embedded cognitive processing.