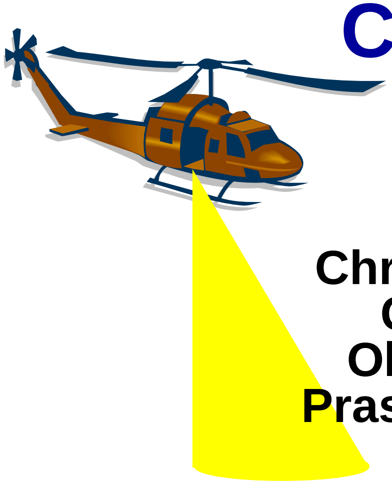
CEARCH

Cognition Enabled ARCHitecture



Stephen Crago and Janice McMahon, USC/ISI

Chris Archer[1], Krste Asanovic[2], Richard Chaung[3], Keith Goolsbey[4], Mary Hall[5], Christos Kozyrakis[6], Kunle Olukotun[6], Una-May O'Reilly[2], Rick Pancoast[7], Viktor Prasanna[8], Rodric Rabbah[2], Steve Ward[2], Donald Yeung[9]

September 20, 2006

[1]Northrop Grumman, [2]MIT, [3]Army I2WD, [4]Cycorp, [5]USC/ISI, [6]Stanford University, [7]Lockheed Martin, [8]USC, [9]University of Maryland

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency (DARPA) or the U.S. Government. Effort sponsored by the Defense Advanced Research Projects Agency (DARPA) through the Department of the Interior National Business Center under grant number NBCH104009.

Outline

- Project Goals
- Architecture Characteristics
- Application Examples
- Summary

Approved for Public Release, Distribution Unlimited

CEARCH Goals

- Develop a *computer architecture* that supports cognitive information processing
  - □ Computer architecture: a set of hardware and system software interfaces and implementations
- Support real-time, embedded cognitive processing requirements through an efficient, high-performance computer architecture
- Identify algorithms and improved algorithm implementations that can leverage the CEARCH computer architecture

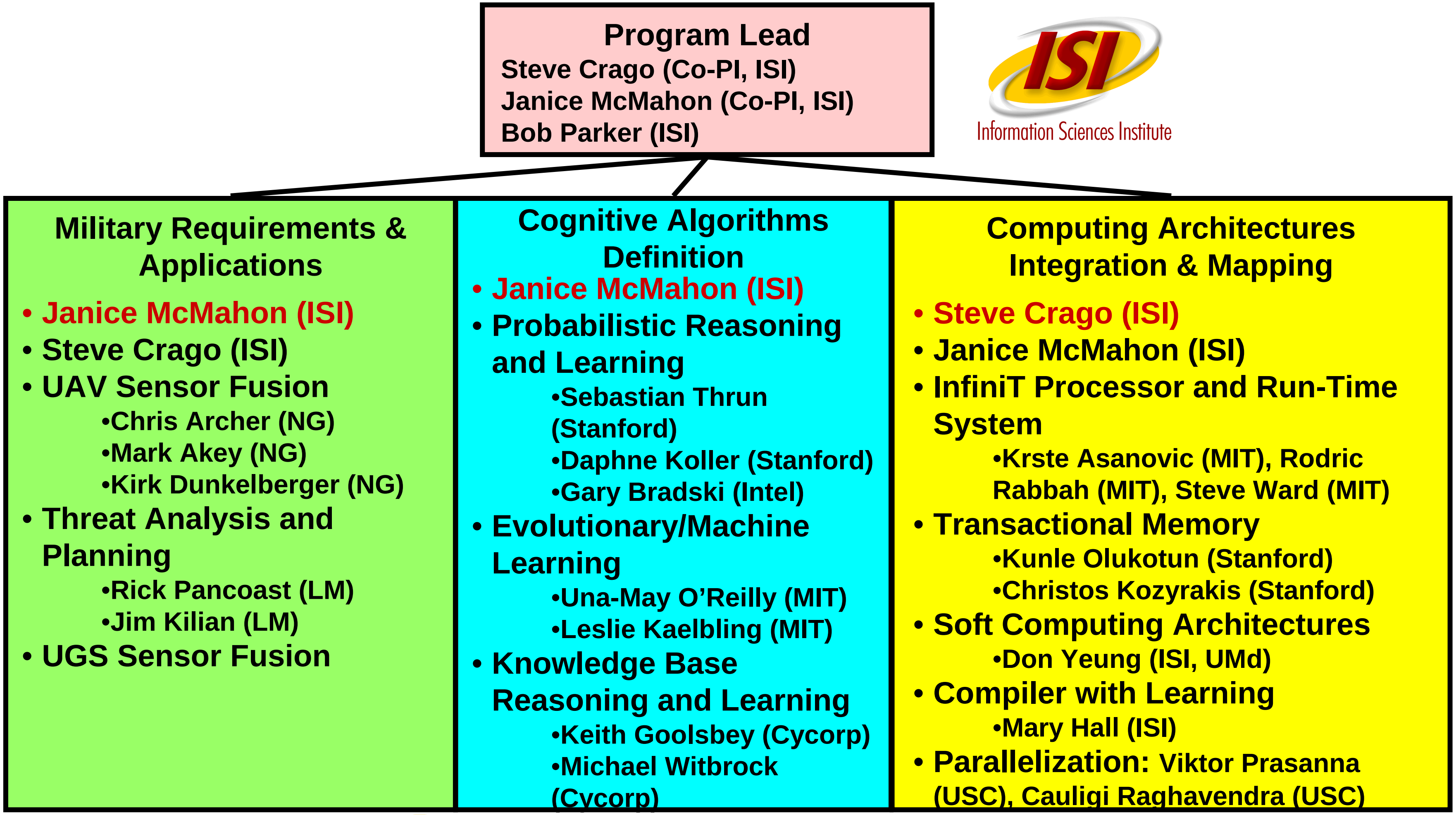
- CEARCH is not a *cognitive architecture* project
  - □ Cognitive architecture: a computational model (usually expressed in software) for a complete cognitive system that may or may not be based on human psychology

CEARCH and Cognitive Architectures

- **The CEARCH computer architecture will run a *variety* of cognitive architectures efficiently**
  - □ **Multiple cognitive architectures important**
    - **No single consensus on cognitive architectures**
    - **Important to support emerging cognitive architecture research: each IPTO program in this domain has its own cognitive architecture**
    - **Different domains may require different cognitive architectures**
  - □ **Support for variety of cognitive architectures**
    - **Wide range of cognitive algorithms drive CEARCH architecture to ensure coverage**
    - **Adaptivity and scalability emphasized to support dynamic processing requirements critical to all cognitive architectures**
- **CEARCH computer architecture has some characteristics of a cognitive system**
  - □ **Introspection and self-management: knows what it is doing and how to process efficiently**
  - □ **Learns how to process more efficiently over time**
  - □ **Supports inexact computations when optimality is not feasible or possible**
  - □ **Robust processing in the context of faults**

Approved for Public Release, Distribution Unlimited

CEARCH Team



Program Lead
Steve Crago (Co-PI, ISI)
Janice McMahon (Co-PI, ISI)
Bob Parker (ISI)

Military Requirements & Applications

- Janice McMahon (ISI)
- Steve Crago (ISI)
- UAV Sensor Fusion
  - •Chris Archer (NG)
  - •Mark Akey (NG)
  - •Kirk Dunkelberger (NG)
- Threat Analysis and Planning
  - •Rick Pancoast (LM)
  - •Jim Kilian (LM)
- UGS Sensor Fusion

Cognitive Algorithms Definition

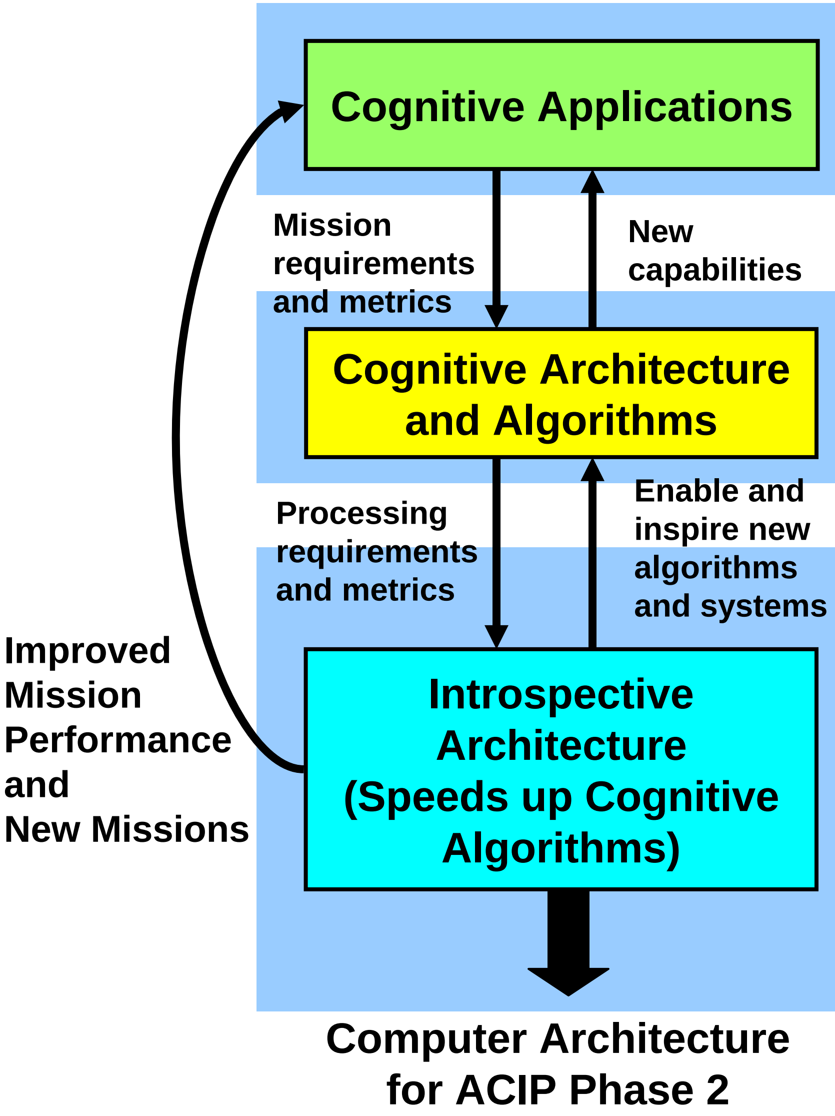
- Janice McMahon (ISI)
- Probabilistic Reasoning and Learning
  - •Sebastian Thrun (Stanford)
  - •Daphne Koller (Stanford)
  - •Gary Bradski (Intel)
- Evolutionary/Machine Learning
  - •Una-May O'Reilly (MIT)
  - •Leslie Kaelbling (MIT)
- Knowledge Base Reasoning and Learning
  - •Keith Goolsbey (Cycorp)
  - •Michael Witbrock (Cycorp)

Computing Architectures Integration & Mapping

- Steve Crago (ISI)
- Janice McMahon (ISI)
- InfiniT Processor and Run-Time System
  - •Krste Asanovic (MIT), Rodric Rabbah (MIT), Steve Ward (MIT)
- Transactional Memory
  - •Kunle Olukotun (Stanford)
  - •Christos Kozyrakis (Stanford)
- Soft Computing Architectures
  - •Don Yeung (ISI, UMd)
- Compiler with Learning
  - •Mary Hall (ISI)
- Parallelization: Viktor Prasanna (USC), Cauligi Raghavendra (USC)

Approved for Public Release, Distribution Unlimited

CEARCH Project Overview



Cognitive Applications

Mission requirements and metrics

New capabilities

Cognitive Architecture and Algorithms

Processing requirements and metrics

Enable and inspire new algorithms and systems

Introspective Architecture (Speeds up Cognitive Algorithms)

Improved Mission Performance and New Missions

Computer Architecture for ACIP Phase 2

- Compact Applications
- DoD SWEPT requirements

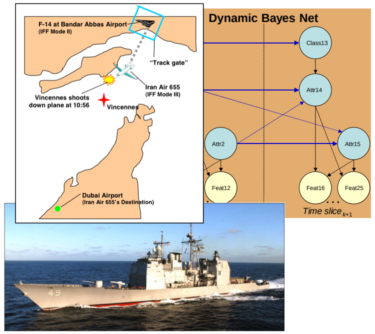
- Probabilistic Reasoning and Learning
- Symbolic Reasoning and Learning
- Planning
- Learning using Evolutionary Algorithms

- Software: Languages and Algorithms
- System Software: Compilers with Learning and Introspective Run-Time
- Hardware: Introspective multi-threading models, Coherence and Consistency, Multi-precision operations, Introspective interconnect

Approved for Public Release, Distribution Unlimited

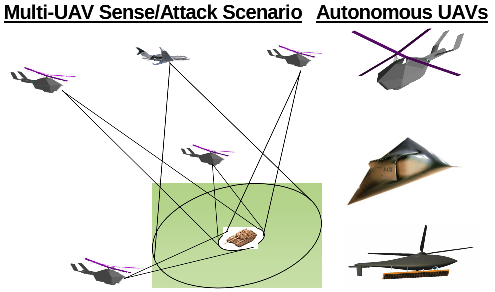
Scenario Summary

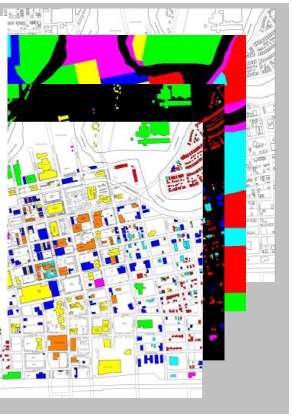
Shipboard Threat Analysis and Planning


F-14 at Bandar Abbas Airport (IFF Mode II)
"Track gate"
Iran Air 655 (IFF Mode III)
Vincennes shoots down plane at 10:56
Vincennes
Dubai Airport (Iran Air 655's Destination)
Dynamic Bayes Net
Class13
Attr14
Attr2
Attr15
Feat12
Feat16
Feat25
Time slice k+1

UGS Urban Situational Awareness

UAV-based Behavior Spotting





Kernel	Example Scenario Requirement	Example architectural drivers
Probabilistic Relational Model (Learn, Infer)	1-2 Tera-updates / sec on large graphs	Probabilistic computation
SATisfiability-based Planner	1 Giga-Boolean-inferences / sec	Parallel tree traversal
Support Vector Machine Classification	2 Tera-ops (variable-precision floating point) / sec	Flexible caching for sparse vectors
Information-form Data Association Tracking	2 Tera-ops (probability calculations) / sec	Parallel sparse matrix calculations
Symbolic Reasoning and Learning	313K problem trees per second	Symbolic matching, irregular memory accesses
System		**Rapid High-Level Reorganization and Responsivity**

Cognitive reasoning and learning techniques require new computing platforms to enable new real-time, embedded capabilities and missions

Must combine orders of magnitude performance/efficiency improvement with ability to respond rapidly to the needs of dynamic environments

Outline

- Project Goals
- Architecture Characteristics
- Application Examples
- Summary

Approved for Public Release, Distribution Unlimited

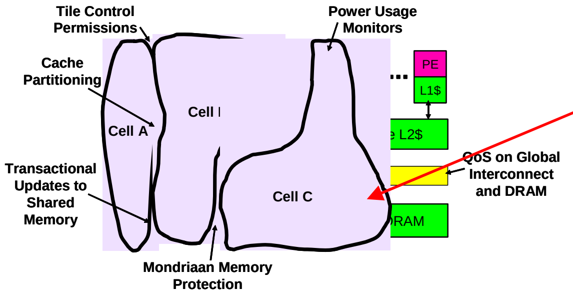
Why Do We Need Hardware for Cognitive Systems?

- **Introspective and Self-Managing Computing**
  - □ **Must support introspective information flow from applications to hardware (and back) to support cognitive resource management and introspective applications**
  - □ **Scalable Web of Cognitive Virtual Processing Elements**
    - **Efficient, high-performance computation required to support real-time reasoning and learning requirements**
    - **Must be adaptable and able to support variety of cognitive processing paradigms (graphs, symbolic reasoning, etc.) and dynamic requirements**
  - □ **Multi-level Soft Computing**
    - **Support for probabilistic and inexact data types and computation pervasive in system (processing, memory, communication, programming model, run-time system)**
  - □ **Adaptive memory system**
    - **Unpredictable, irregular memory accesses and large working sets**
    - **Driven by parallel computation, dynamic resource allocation, and fundamental characteristics of algorithms and data**

Approved for Public Release, Distribution Unlimited

Introspection and Self-Management

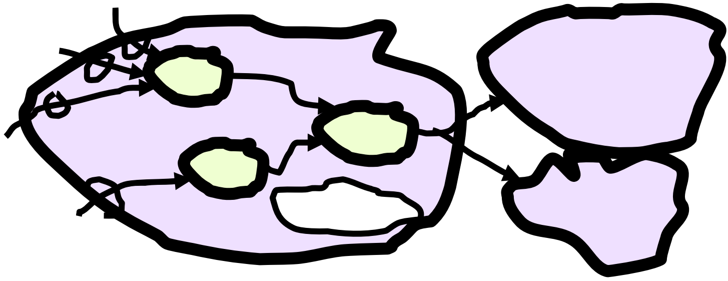
- System must adapt to unpredictability in cognitive systems
  - □ Dynamic scenarios lead to dynamic and unpredictable changes in processing requirements
  - □ Cognitive processing too complex to be managed by programmer
    - Cognitive algorithms provide means for system to manage itself
  - □ Faults are unavoidable at this scale
- Introspection required to support autonomous adaptability
  - □ **Processing:** precision, performance required, operation mixes, efficiency of functional units
  - □ **Memory and Communication:** access/communication patterns, cache hit rates, working set sizes, precision required, bandwidth/latency trade-offs, protection

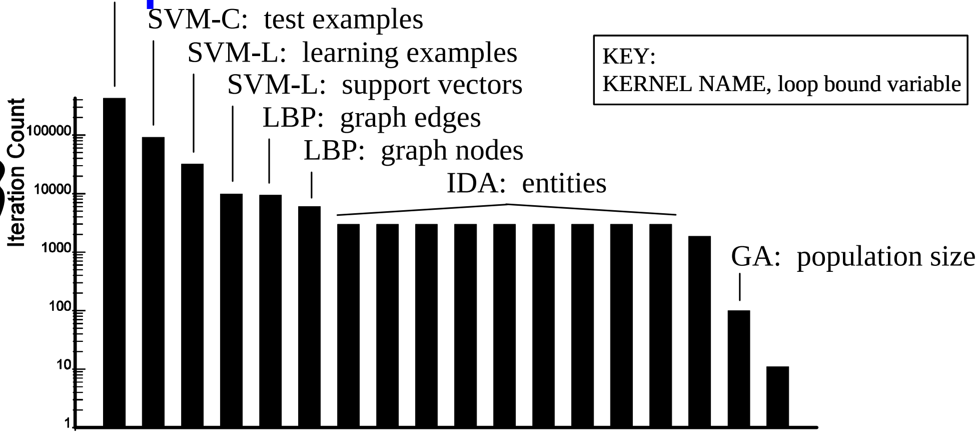

Tile Control Permissions
Power Usage Monitors
Cache Partitioning
Cell A
Cell I
Cell C
PE
L1$
e L2$
QoS on Global Interconnect and DRAM
RAM
Transactional Updates to Shared Memory
Mondriaan Memory Protection

Cell-based introspection and management

Scalable Web of Cognitive Virtual Processing Elements

- Cognitive processing requires massive fine-grained parallelism with highly efficient processing elements
- Cognitive processing elements different from general-purpose computing, scientific computing, and signal processing elements
  - □ Processing granularity highly variable and dynamic
  - □ Cognitive systems and scenarios lead to dynamic code and data movement and load balancing
  - □ Density of parallelism must be much higher to do real-time reasoning and learning in complex scenarios



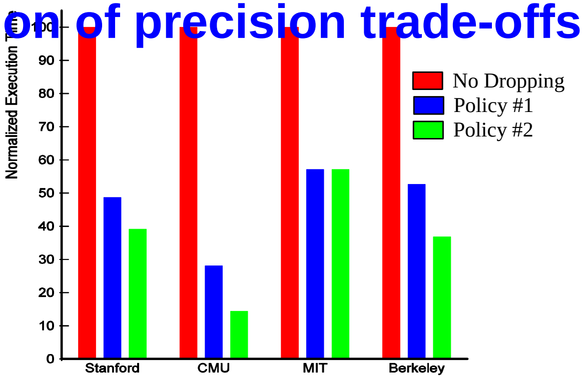


Parallelism With Varying Granularity and Computation Types

Multi-Level Soft Computing

- Exploit the tolerance for imprecision, uncertainty, partial truth, and approximation to achieve tractability, robustness and low solution cost*
  - □ Optimality or exactness infeasible in cognitive application domains
  - □ Input data has imprecision and inaccuracy
  - □ Robustness needed to handle transient and persistent faults
- Exploitation of soft computing for performance gains changes architecture at all levels
  - □ Processor: data types, functional units, circuit design
  - □ Memory: local and shared lossy memory protocols, latency reduction
  - □ Communication: lossy protocols, QoS tuning
  - □ System software: data types, communication of precision trade-offs to programmer, resource management

Performance Improvements From Message Dropping



*http://www.soft-computing.de/def.html

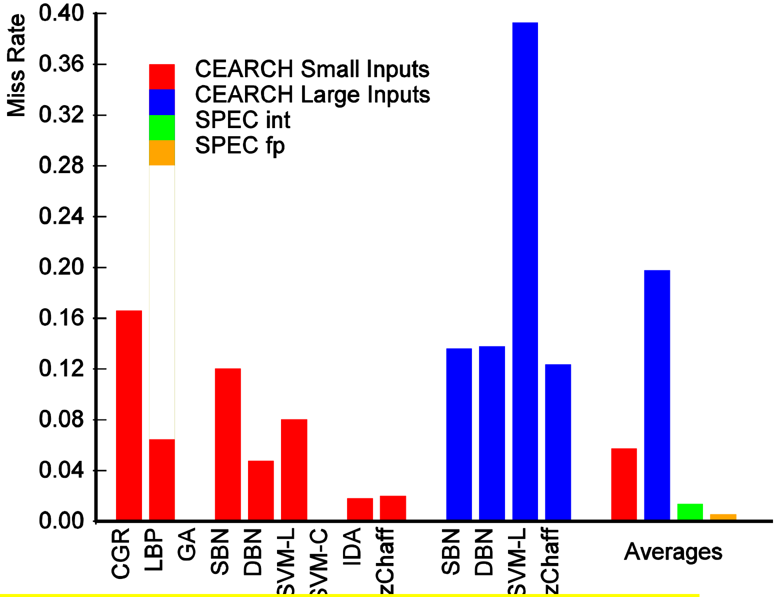
Approved for Public Release, Distribution Unlimited

Adaptive Memory System

- **Cognitive processing leads to poor memory system behavior in traditional memory systems**
  - □ Some algorithms have irregular and hard-to-predict access patterns
  - □ Working sets can be very large because of complexity of scenarios
  - □ Dynamic resource allocation and fine-grained parallelism leads to more global memory accesses and locality challenges

- **Memory system requirements**
  - □ Flexible allocation among cognitive processing elements
  - □ Fine-grained protection
  - □ Flexible commit policies
  - □ Inexpensive roll-back for fault tolerance and race conditions between parallel compute elements

L1 Cache



Miss Rates for Cognitive Algorithms Using Traditional Cache

CEARCH Architecture Layers

Programming Model

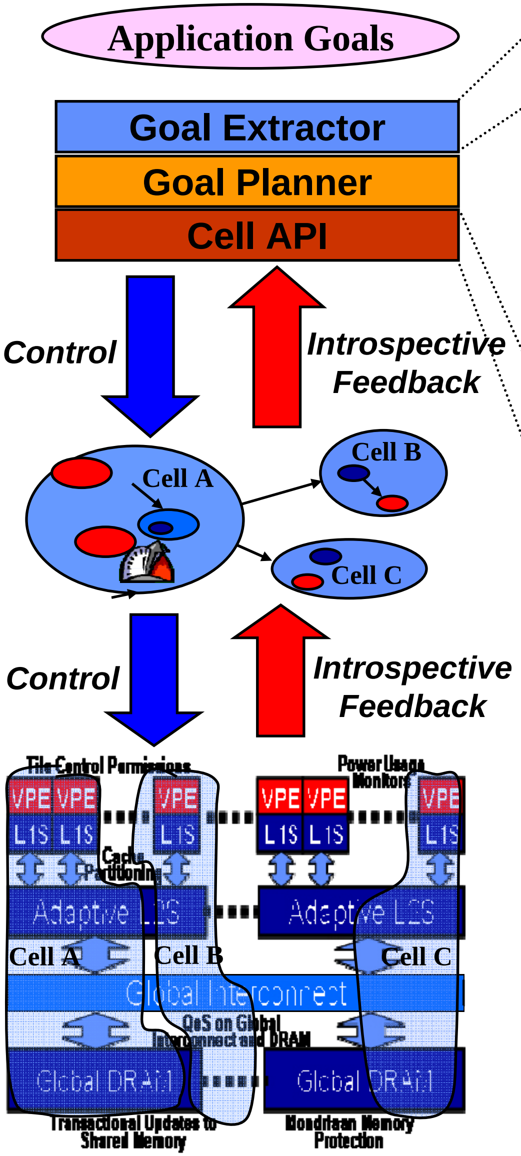
- Abstraction barriers provide scalable low-level performance with high-level specifications
- Goal-based performance and resource allocation allows computation to be in part selected by system
- Soft computing semantics

Runtime System

- Learning and reasoning-based goal-oriented instrumentation and compilation
- Adaptive and introspective hierarchical resource allocation for processing, memory, and communication

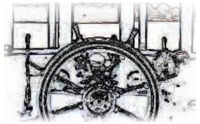
Hardware Architecture

- Millions of introspective virtual processing elements running on thousands of hardware engines
- Adaptive memory for efficient data access and sharing
- Soft computing support



Programming Model for the Algorithm

- “The Bridge”



  - □ Language expresses the algorithm and algorithm goals
  - □ Architecture independent and malleable code

Programming Model for Introspection

- “The Engine Room”

  - □ Can analyze the program (“reflection” interface)
  - □ Can find information about the resources/architecture
  - □ Provide rules for
    - Scheduling and Resource allocation
    - Learning and Adaptation
    - Soft computing and fault tolerance
  - □ By
    - Default policies
    - Overwritten by creating generic rules
    - Or custom rules for an application

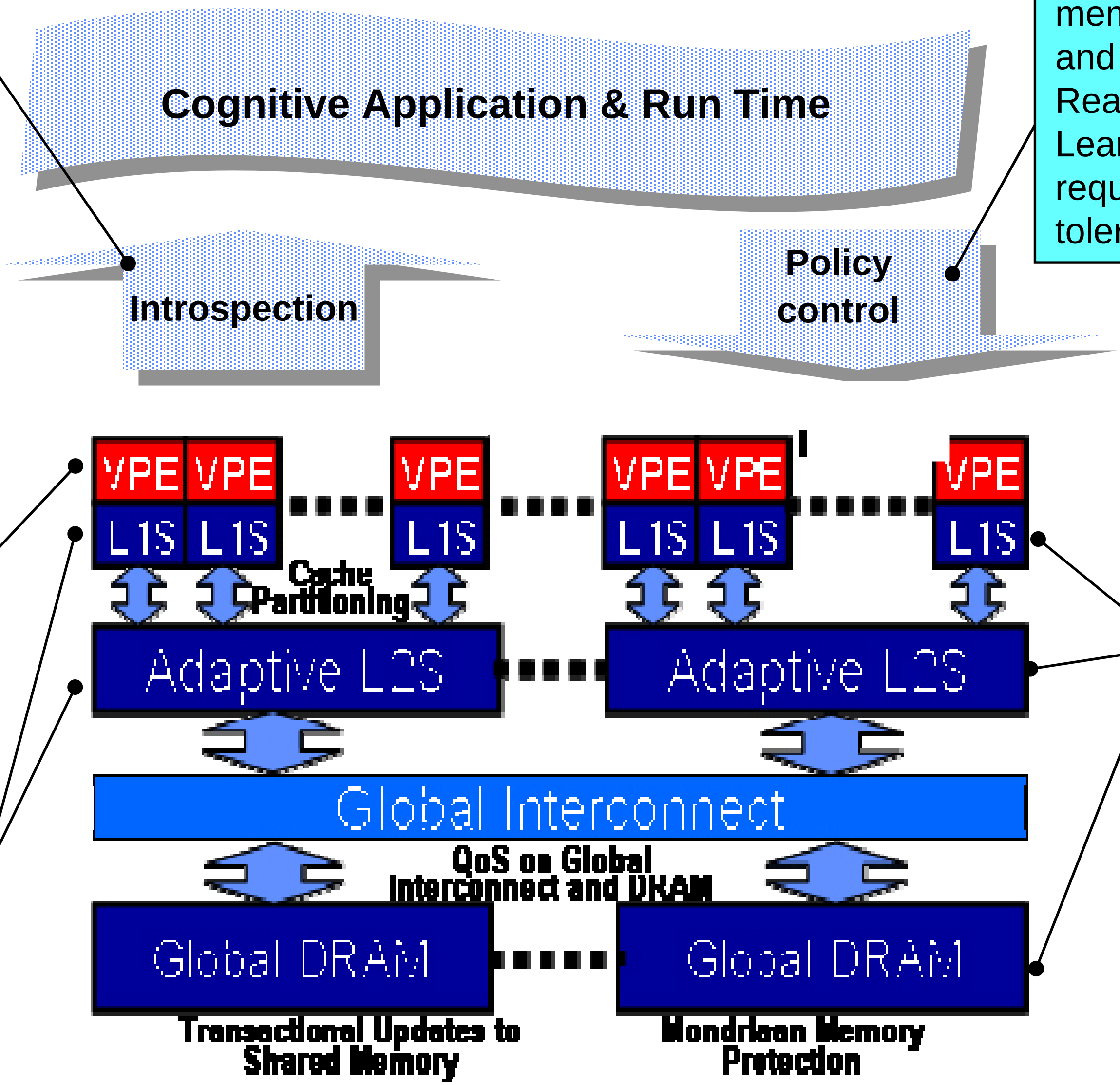
Approved for Public Release, Distribution Unlimited

CEARCH Hardware Architecture

Performance, Communication, Resource availability, Failure, Power

Stored processor *Millions* of scalable cognitive virtual processing elements (***stored threads)*** for dynamic parallel *reasoning* and *learning*, ***Soft computing***

Multi-level cognitive memory, stored processor working sets



Processor and memory allocation and precision, Reasoning and Learning requirements, Fault tolerance

Adaptive transactional Mondriaan memory, ***Parallel reasoning and learning data accesses***, ***Soft coherence***, ***Speculation,*** Locality management, ***Cell*** Sharing, Isolation and Protection

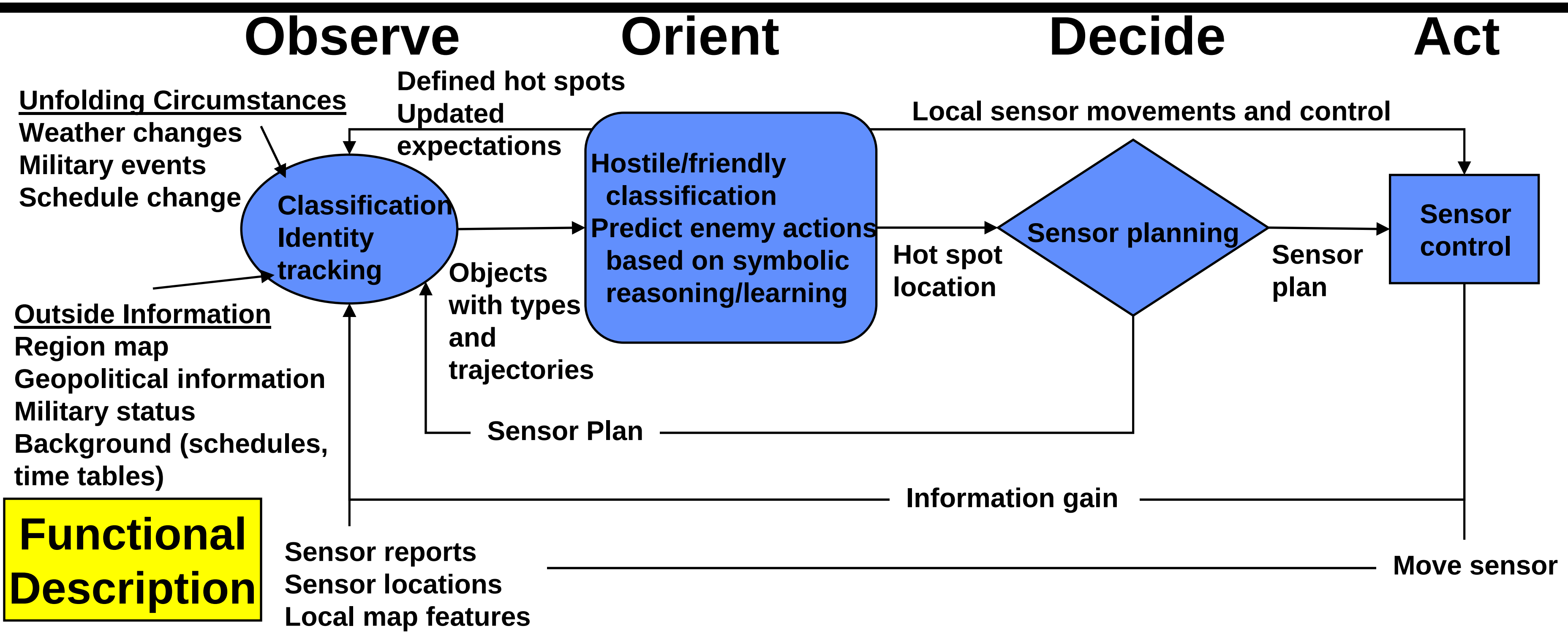
Approved for Public Release, Distribution Unlimited

Outline

- Project Goals
- Architecture Characteristics
- Application Examples
- Summary

Approved for Public Release, Distribution Unlimited

Spotting Behaviors OODA Loop



Observe

Orient

Decide

Act

Unfolding Circumstances
Weather changes
Military events
Schedule change

Defined hot spots
Updated
expectations

Local sensor movements and control

Classification
Identity
tracking

Hostile/friendly
classification
Predict enemy actions
based on symbolic
reasoning/learning

Sensor planning

Sensor
control

Objects
with types
and
trajectories

Hot spot
location

Sensor
plan

Outside Information
Region map
Geopolitical information
Military status
Background (schedules,
time tables)

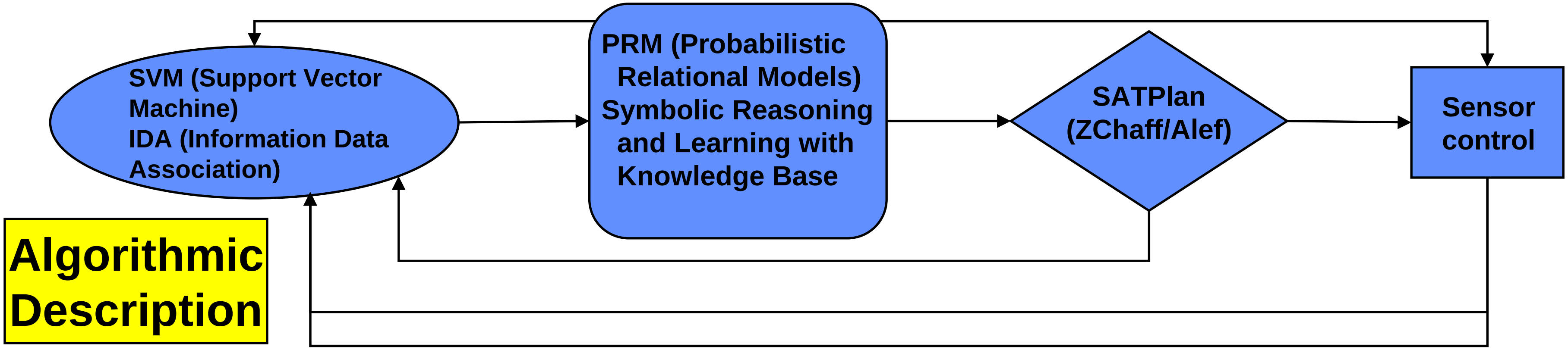
Sensor Plan

Information gain

Functional
Description

Sensor reports
Sensor locations
Local map features

Move sensor



SVM (Support Vector
Machine)
IDA (Information Data
Association)

PRM (Probabilistic
Relational Models)
Symbolic Reasoning
and Learning with
Knowledge Base

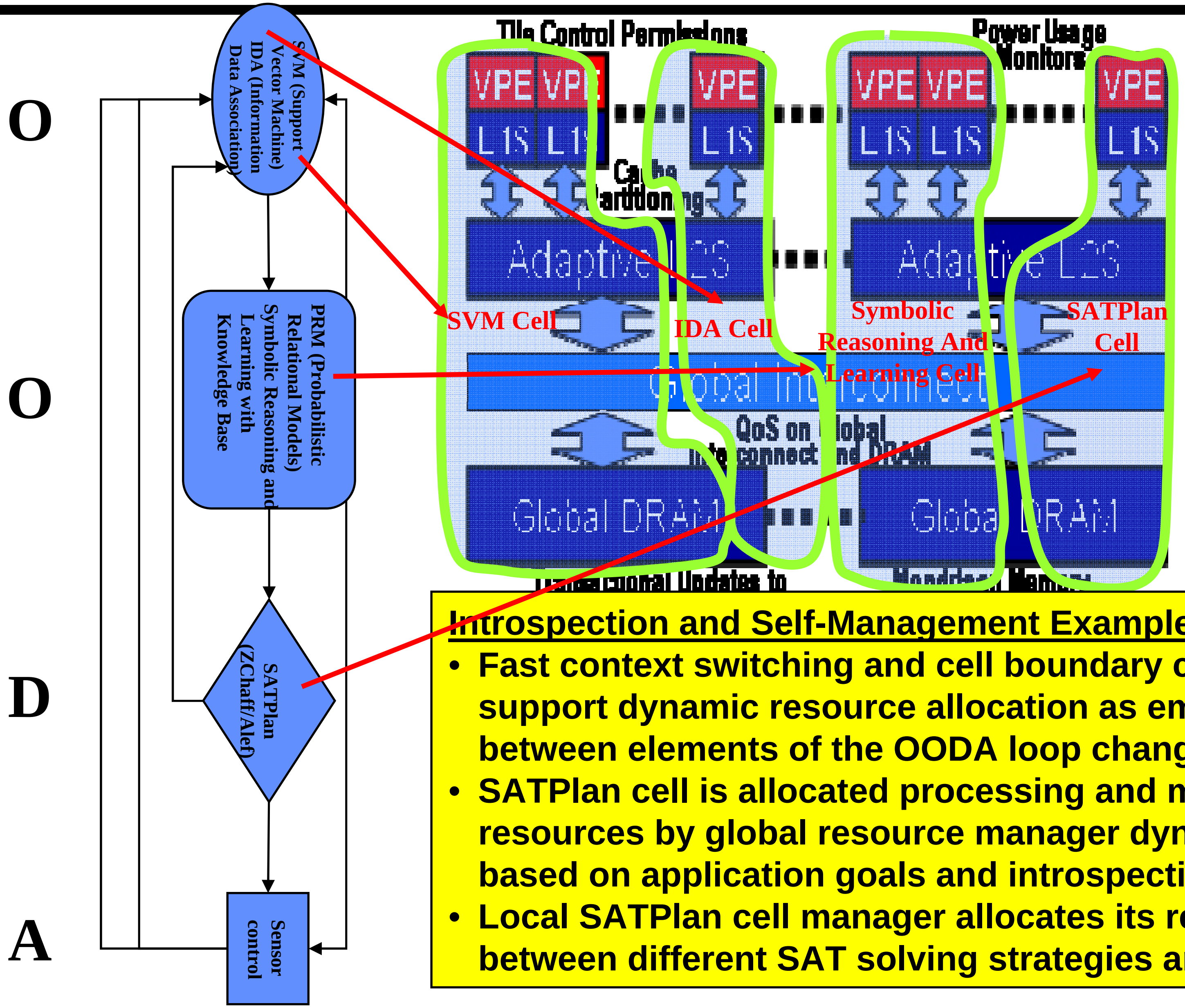
SATPlan
(ZChaff/Alef)

Sensor
control

Algorithmic
Description

Approved for Public Release, Distribution Unlimited

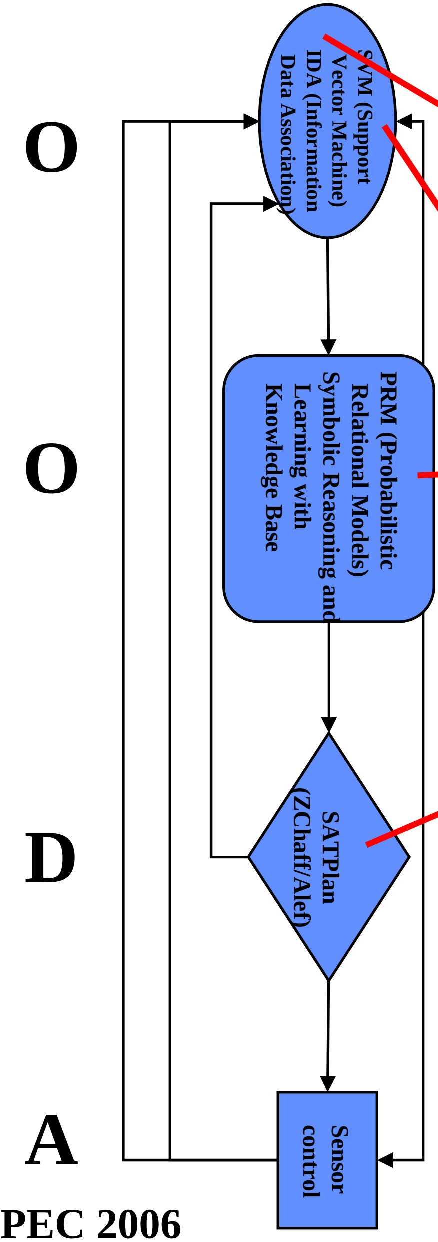
Introspection and Self-Management

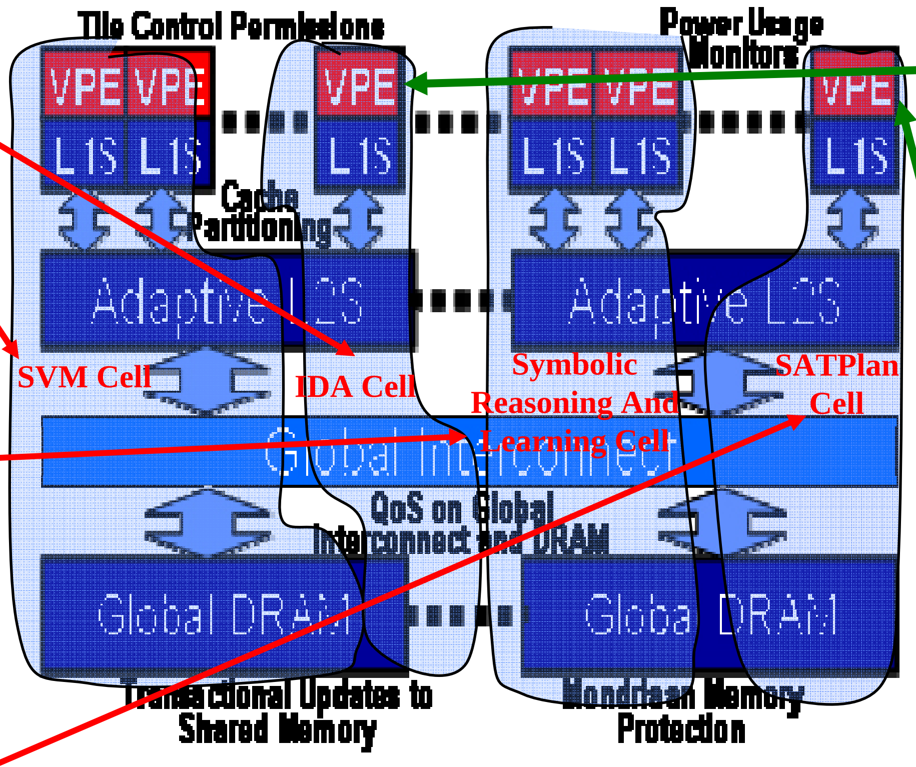

O
O
D
A
SVM (Support Vector Machine)
IDA (Information Data Association)
PRM (Probabilistic Relational Models)
Symbolic Reasoning and Learning with Knowledge Base
SATPlan (ZChaff/Alef)
Sensor control
Tile Control Permissions
Power Usage Monitors
VPE
L1S
Cache Partitioning
Adaptive L2S
Global Interconnect
QoS on Global Interconnect and DRAM
Global DRAM
Transactional Updates to Redundant Memory
SVM Cell
IDA Cell
Symbolic Reasoning And Learning Cell
SATPlan Cell

Introspection and Self-Management Examples

- Fast context switching and cell boundary changes support dynamic resource allocation as emphasis between elements of the OODA loop changes
- SATPlan cell is allocated processing and memory resources by global resource manager dynamically based on application goals and introspective monitors
- Local SATPlan cell manager allocates its resources between different SAT solving strategies and sub-goals

Scalable Web of Virtual Processing Elements


O
O
D
A
SVM (Support Vector Machine)
IDA (Information Data Association)
PRM (Probabilistic Relational Models)
Symbolic Reasoning and Learning with Knowledge Base
SATPlan (ZChaff/Alef)
Sensor control



Virtual processing element configured for probabilistic reasoning and learning

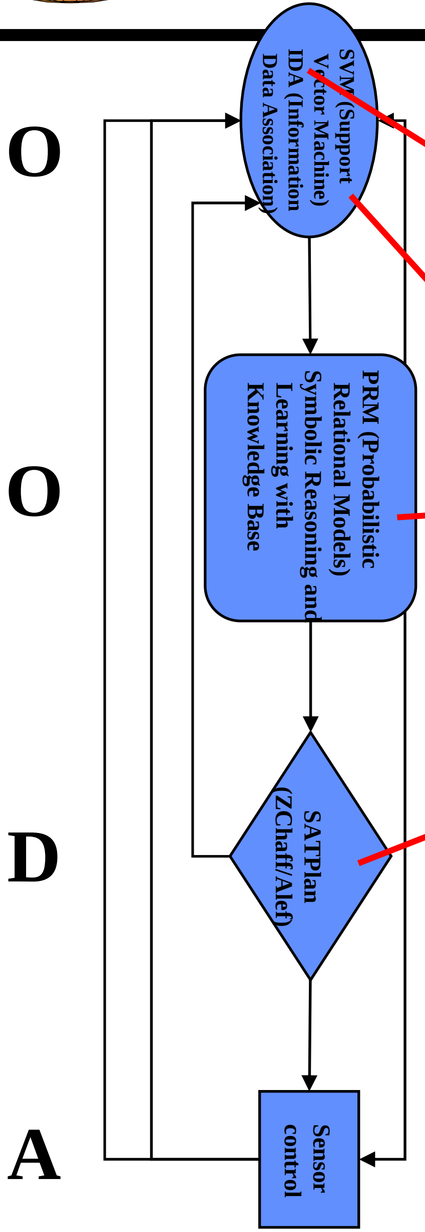
Virtual processing element configured for logical reasoning and learning

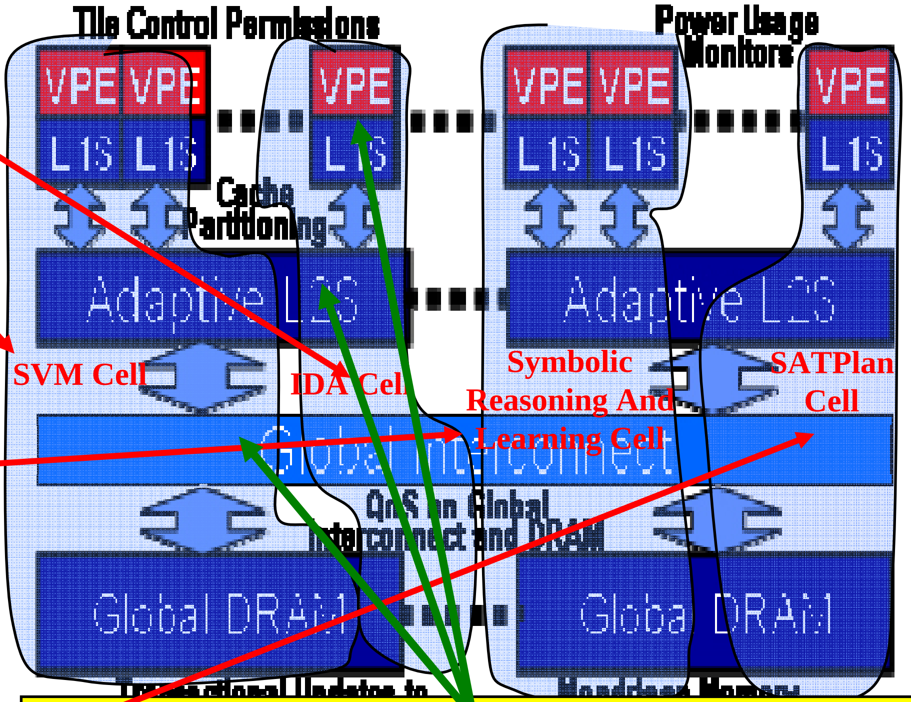
Scalable Web of Virtual Processing Elements

- **Millions of threads implement graph nodes in SVM, IDA, symbolic reasoning and learning, and SATPlan and are available for introspection in memory; thousands are active at given time**
- **Processing elements are configured for computation type (e.g. probabilistic for IDA and logical for SATPlan)**

Approved for Public Release, Distribution Unlimited

Multi-Level Soft Computing


O
SVM (Support Vector Machine)
IDA (Information Data Association)
O
PRM (Probabilistic Relational Models)
Symbolic Reasoning and Learning with Knowledge Base
D
SATPlan (ZChaff/Alef)
A
Sensor control

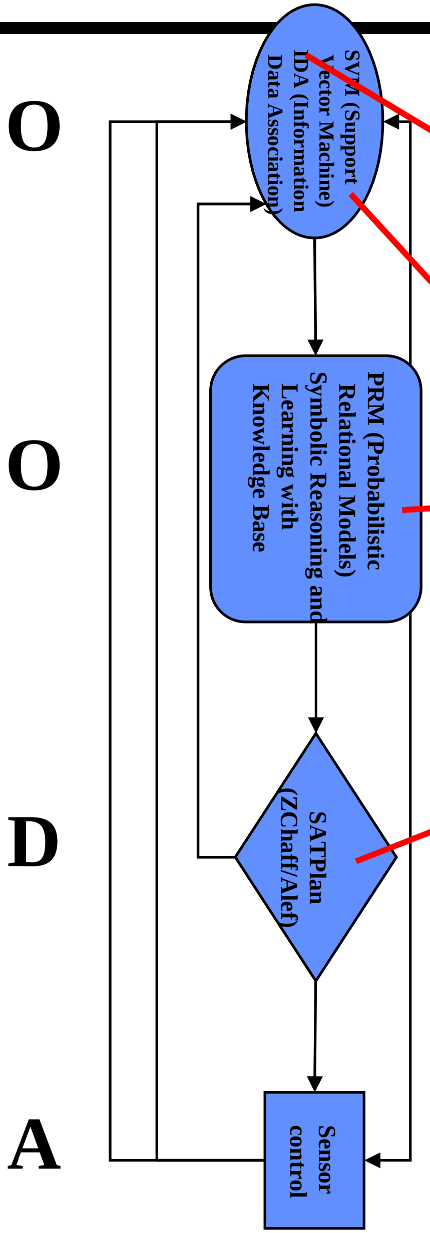


Multi-Level Soft Computing

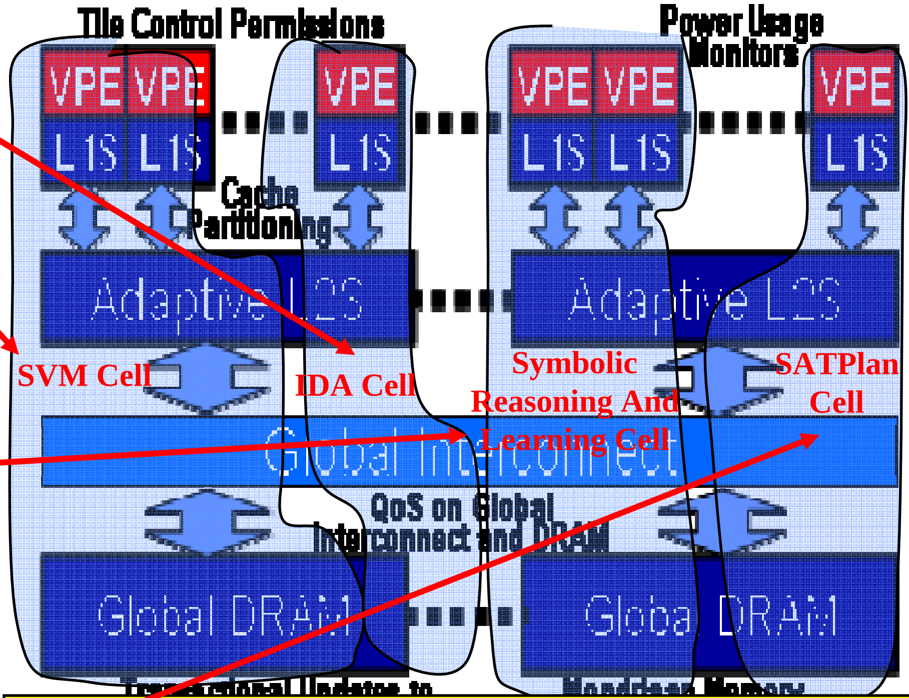
- Algorithms can change computation to match accuracy requirements (e.g. change IDA update frequency)
- Processing elements compute probabilistic data for IDA with variable-precision
- Soft commits in transactional memory can allow conflicting updates in probabilistic IDA computations
- QoS in interconnect can be varied for messages from approximate algorithms

Approved for Public Release, Distribution Unlimited

Adaptive Memory



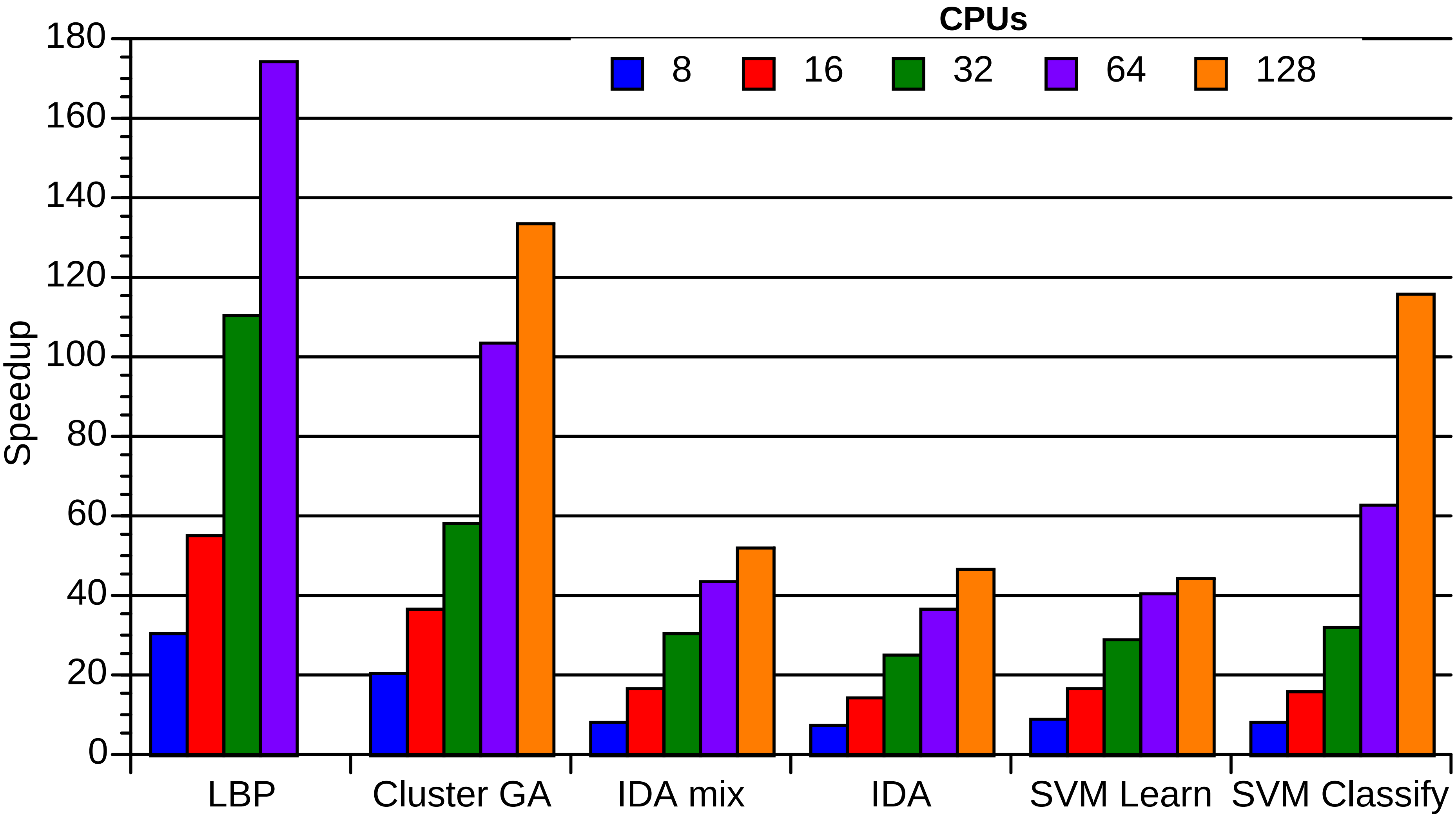
O
O
D
A
SVM (Support Vector Machine)
IDA (Information Data Association)
PRM (Probabilistic Relational Models)
Symbolic Reasoning and Learning with Knowledge Base
SATPlan (ZChaff/Alef)
Sensor control



Adaptive Memory

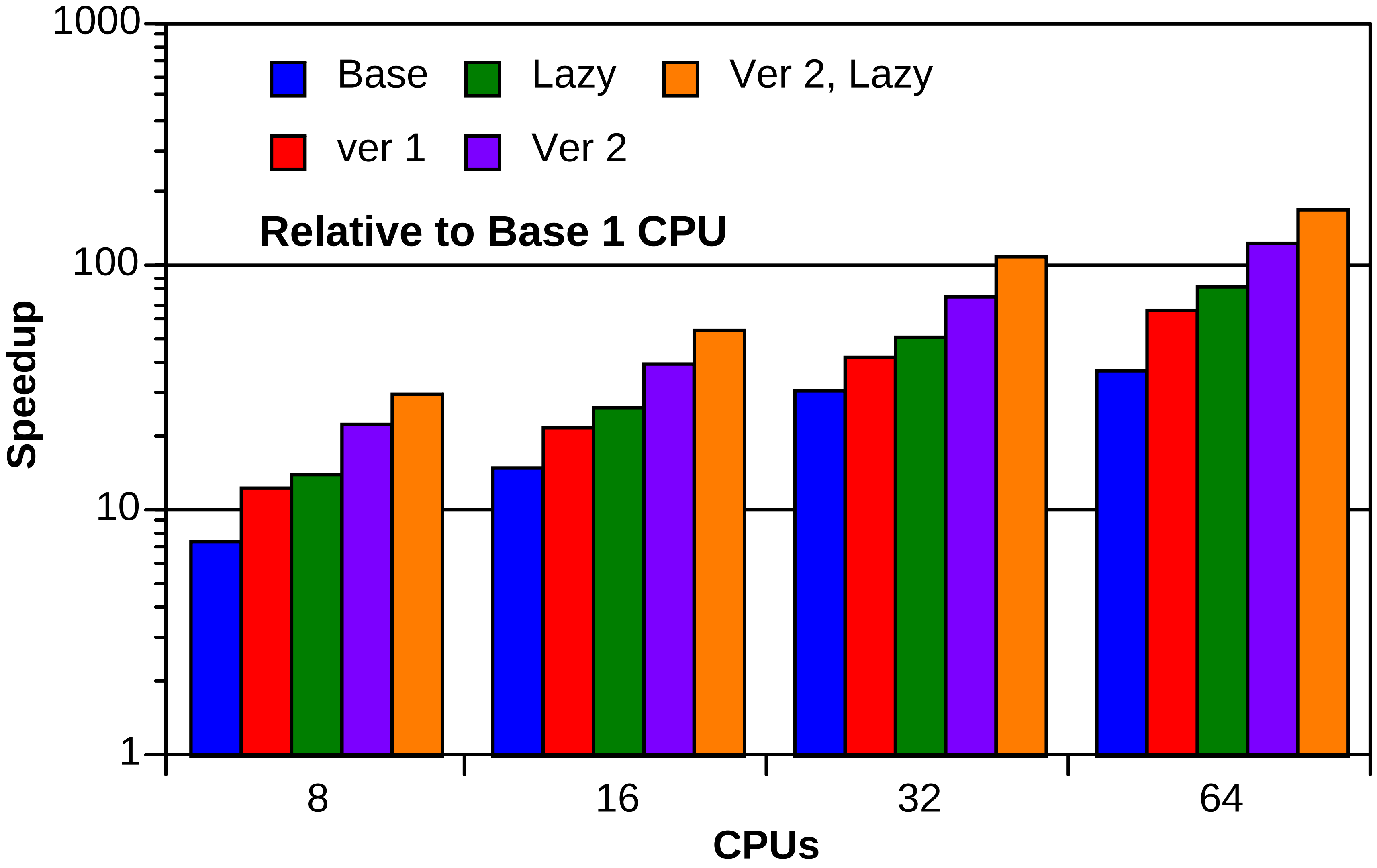
- L2 and global memory can be dynamically reallocated between cells to match working sets (e.g. SVM may have a larger working set than IDA)
- Symbolic reasoning may require a stricter transaction coherency protocol than probabilistic reasoning (IDA)
- Mondriaan memory protection allows fine-grain sharing between SVM and IDA and also between sub-goals of symbolic reasoning and planning

CEARCH Application Speedups



Approved for Public Release, Distribution Unlimited

LBP Performance Improvement



Approved for Public Release, Distribution Unlimited

Outline

- Project Goals
- Architecture Characteristics
- Application Examples
- Summary

Approved for Public Release, Distribution Unlimited

CEARCH Summary

- CEARCH is a dynamic self-managing architecture for cognitive processing uniquely suited to complex environments
  - □ Driven by cognitive system and algorithm characteristics
  - □ Dynamically organize resources to optimize performance, power and reliability
  - □ Adaptation and introspection in both hardware and software
- CEARCH has unique features to efficiently support cognitive applications and that provide capability not possible with today's COTS architectures
  - □ Stored processor
  - □ Adaptive, transactional memory
  - □ Soft computation
  - □ Introspection and run-time policy control support
- Preliminary architecture evaluation indicates
  - □ High performance potential
  - □ Well suited to cognitive applications and soft computing